Improving Multilingual Neural Machine Translation with Auxiliary Source Languages

Weijia Xu\textsuperscript{1} Yuwei Yin\textsuperscript{2} Shuming Ma\textsuperscript{3} Dongdong Zhang\textsuperscript{3} Haoyang Huang\textsuperscript{3}
\textsuperscript{1}Department of Computer Science, University of Maryland
\textsuperscript{2}State Key Lab of Software Development Environment, Beihang University
\textsuperscript{3}Microsoft Research Asia
weijia@cs.umd.edu; yuweiyin@buaa.edu.cn; 
{shumma, dozhang, haohua}@microsoft.com

Abstract

Multilingual neural machine translation models typically handle one source language at a time. However, prior work has shown that translating from multiple source languages improves translation quality. Different from existing approaches on multi-source translation that are limited to the test scenario where parallel source sentences from multiple languages are available at inference time, we propose to improve multilingual translation in a more common scenario by exploiting synthetic source sentences from auxiliary languages. We train our model on synthetic multi-source corpora and apply random masking to enable flexible inference with single-source or bi-source inputs. Extensive experiments on Chinese/English→Japanese and a large-scale multilingual translation benchmark show that our model outperforms the multilingual baseline significantly by up to +4.0 BLEU with the largest improvements on low-resource or distant language pairs.

1 Introduction

Neural machine translation (NMT) has achieved the state-of-the-art performance across domains and language pairs (Wu et al., 2016; Bojar et al., 2018; Hassan et al., 2018; Barrault et al., 2019). One of the advantages of NMT over statistical machine translation models is that it enables information sharing among high-resource and low-resource languages by training a multilingual model on the parallel data from multiple language pairs, which has been shown to improve translation quality, especially on low-resource language pairs (Firat et al., 2016a; Ha et al., 2016; Aharoni et al., 2019).

Although multilingual NMT models typically handle one language pair at a time during both training and inference (Ha et al., 2016; Johnson et al., 2017), prior work has shown that translating from multiple parallel source sentences can further improve translation quality (Och and Ney, 2001; Zoph and Knight, 2016; Garmash and Monz, 2016; Nishimura et al., 2018). They propose multi-source translation models to exploit multiple source inputs at inference time. However, these models are limited to the application scenario where the source sentence has already been manually translated into multiple languages. We argue that, in the more common scenario where only one source sentence is provided, we could also improve the translation quality of multilingual NMT models by augmenting the source input with a synthetic sentence generated by a translation model into another language. As shown by the example in Figure 1, the additional synthetic sentence can help translate low-frequency and domain-specific words that are difficult to translate directly from the source.

In this paper, we propose a novel bi-source multilingual NMT model that leverages a synthetic source sentence from an auxiliary language to better translate a source sentence into the target language. We train our bi-source NMT model on a synthetic multi-source translation corpus generated by translating the source side of the parallel data
into other source languages using pre-trained NMT models. We contribute a novel training algorithm that 1) randomly selects the auxiliary language at each training iteration, which improves the multilinguality of the encoder representations, and 2) randomly masks out the auxiliary sentence during training, so that the model can perform inference flexibly in two different modes, including a) single-source inference where our model takes a single source as input, and b) bi-source inference where we first translate the original source to another language using an NMT model and then feed the two source sentences into our model to predict the target translation. This allows end users to balance between translation quality and latency by choosing different inference modes.

We experiment on the ASPEC Chinese and English-to-Japanese translation and a large-scale English-to-many translation benchmark that includes 10 language pairs from WMT. Results show that our method is simple yet effective – it improves English→Japanese translation on out-of-domain test sets and outperforms strong baselines by an average of +1.9 BLEU on the English-to-many translation benchmark. The largest improvements are on low-resource languages, where it brings up to +4.0 BLEU improvements. Further analysis confirms our hypothesis that bi-source inference helps the model disambiguate word senses during translation.

## 2 Bi-Source Multilingual NMT

Inspired by prior work on multi-source translation (Zoph and Knight, 2016; Nishimura et al., 2018), we hypothesize that multilingual translation models can benefit from additional synthetic source sentences that are automatically translated from the original source.

### 2.1 Model

Formally, the model computes the probability of target sentence $y^t_i$ in language $l_t$ given the original source sentences $x^{ls}_{l_s}$ and a synthetic source sentence $\tilde{x}^{la}_{l_a}$ translated from $x^{ls}_{l_s}$ into an auxiliary language $l_a$ ($l_a \neq l_s$, $l_a \neq l_t$) by an MT model:

$$p(y^t_i | x^{ls}_{l_s}, \tilde{x}^{la}_{l_a}) = p(y^t_i | f(x^{ls}_{l_s}, \tilde{x}^{la}_{l_a}; \Theta_{enc}); \Theta_{dec})$$

(1)

where $\Theta_{enc}$ and $\Theta_{dec}$ represent the encoder and decoder parameters, respectively, and $f(\cdot; \Theta_{enc})$ produces the encoder representations of the inputs.

Our encoder-decoder model is based on the Transformer architecture (Vaswani et al., 2017). As shown in Figure 2, we adopt techniques from context-aware machine translation (Voita et al., 2018) to integrate the additional source input into the model:

### Multi-Encoder Approach

encodes the source sentences using separate encoders (Voita et al., 2018) to obtain the hidden representations $f(x^{ls}_{l_s}; \Theta_{enc}) = H^s$ and $f(\tilde{x}^{la}_{l_a}; \Theta_{enc}) = H^a$. Then, the decoder can attend to $H^s$ and $H^a$ separately and apply a gating mechanism to obtain the fusion vector $h_i$:

$$
\begin{align*}
    h_i^s &= \text{Attn}(H^s, h_i^{tgt}) \\
    h_i^a &= \text{Attn}(H^a, h_i^{tgt}) \\
    g_i &= \sigma(W_q[h_i^s; h_i^a] + b_q) \\
    h_i &= g_i \odot h_i^s + (1 - g_i) \odot h_i^a
\end{align*}
$$

(2)

where $h_i^{tgt}$ represents the hidden state of the $i$-th target token, $W_q$ and $b_q$ are model parameters, and $\sigma$ represents the logistic sigmoid function.

### Single-Encoder Approach

encodes the source sentences by concatenating them into a long sequence (Dabre et al., 2017; Tiedemann and Scherrer, 2017), which is then fed to an embedding layer and a stack of self-attention and position-wise feed-forward layers to produce a sequence of hidden representations $f([x^{ls}_{l_s}; \tilde{x}^{la}_{l_a}]; \Theta_{enc}) = H$. Then, we apply the encoder-decoder attention to the full sequence of encoder representations $H$:

$$h_i = \text{Attn}(H, h_i^{tgt})$$

(3)

The single-encoder approach is simpler than the multi-encoder one and can be easily adapted to multiple auxiliary languages as inputs.

### 2.2 Training

Our bi-source multilingual model is trained on a combination of datasets $\mathcal{D} = \bigcup_{l_0 \in \mathcal{S}, l_0 \in \mathcal{A}, l_t \in \mathcal{T}} \mathcal{D}^{l_0 \times l_0 \times l_t}$, where $\mathcal{S}$ is the set of source languages, $\mathcal{T}$ is the set of target languages, $\mathcal{A}$ represents the set of auxiliary languages, and $\mathcal{D}^{l_0 \times l_0 \times l_t} = \{(x^{ls}_{l_s}, \tilde{x}^{la}_{l_a}, y^t_i)\}$ is a bi-source translation dataset which can be formed by data augmentation via MT. The objective is to
maximize the log-likelihood of the target sentences given the original and auxiliary source sentences:

\[
\mathcal{L} = \sum_{(x^s, \tilde{x}^a, y^t) \in D} \left[ p_{\text{mask}} \log p(y^t | x^s) + (1 - p_{\text{mask}}) \log p(y^t | \tilde{x}^a) \right]
\] (4)

At each training iteration, we randomly pick a triplet of mutually distinct source, auxiliary, and target languages \((l_s, l_a, l_t)\). Next, we randomly sample a batch of training examples \(\{(x^{l_s}, \tilde{x}^{l_a}, y^{l_t})\}\) from \(D_{l_s \times l_a \times l_t}\) and maximize the log probability of the target sentence \(y^{l_t}\) given source sentence \(x^{l_s}\) and auxiliary sentence \(\tilde{x}^{l_a}\). To enable more flexible decoding and to improve model robustness, we randomly mask out the auxiliary sentences with probability \(p_{\text{mask}}\) during training.²

**Creating Pseudo Training Data** We adopt data augmentation techniques (Sennrich et al., 2016a; Nishimura et al., 2018) to construct the bi-source data using parallel data from multiple language pairs. More specifically, we first train a multilingual NMT model \(M_{S \rightarrow A}\) to translate between source and auxiliary languages. Next, we extend each parallel dataset \(\{(x^{l_s}, y^{l_t})\}\) to pseudo bi-source datasets \(\{(x^{l_s}, \tilde{x}^{l_a}, y^{l_t})\}\) by translating \(x^{l_s}\) into auxiliary languages \(l_a\) using \(M_{S \rightarrow A}\). Finally, we combine all pseudo bi-source datasets into the training data \(D\) to train our bi-source model.

**2.3 Inference**

Prior work on multi-source NMT (Zoph and Knight, 2016; Nishimura et al., 2018) assumes access to multi-source inputs at inference time, which has limited their scope of application in the real world. Instead, we test our model in a more realistic scenario where only a single source sentence for each test instance is provided. We experiment with two inference modes: 1) **single-source inference** where we provide our model with only a single source sentence during inference. 2) **bi-source inference** where we first augment the source sentence by translating it into an auxiliary language using the NMT model \(M_{S \rightarrow A}\) and then use our bi-source model to generate the target translation given the original and auxiliary source sentences.

**3 Experiments**

**3.1 Data**

We evaluate our approach on two translation tasks, including Chinese/English→Japanese (Zh/En→Ja) and a large-scale En→X task that translates from English to 10 languages, including French (Fr), Czech (Cs), German (De), Finnish (Fi), Lat...
vian (Lv), Estonian (Et), Romanian (Ro), Hindi (Hi), Turkish (Tr), and Gujarati (Gu).

**Zh/En→Ja** We set the source and auxiliary language sets $S = \{\text{Zh, En}\}$, and the target language set $T = \{\text{Ja}\}$. The training data consists of 0.67M sentence pairs for Japanese-Chinese and 3.0M sentence pairs for Japanese-English from ASPEC corpus (Nakazawa et al., 2016). We use the provided development set and test the models on both in-domain test set from ASPEC and out-of-domain test sets as shown in Table 1. To train the Chinese→English translation model for data augmentation, we use the training corpora (21.2M) from WMT18 (Bojar et al., 2018), newstest2017 as development set, and newstest2018 as test set.

**En→X** We follow the preprocessing steps in Wang et al. (2020): we remove duplicated sentence pairs and the pairs with the same source and target sequences from the training corpora and then tokenize all data using SentencePiece (Kudo and Richardson, 2018) with a shared vocabulary size of 64K tokens. Table 2 shows the training data size after preprocessing and the test set for each language pair.

### 3.3 Training

We use the Transformer models (Vaswani et al., 2017) implemented in fairseq.

**Zh/En→Ja** We use the Transformer base architecture with $d_{\text{model}} = 512$, $d_{\text{hidden}} = 2048$, $n_{\text{heads}} = 8$, $n_{\text{layers}} = 6$, and $p_{\text{dropout}} = 0.1$. We apply label smoothing of 0.1. We adopt Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.0005, batch size of 48,000 tokens, and 4,000 warm-up updates for maximum 500,000 steps or 50 epochs. We select the best checkpoint based on validation perplexity. During inference, we use beam search with a beam size of 8 and length penalty of 1.0.

**En→X** We use the Transformer big model with $d_{\text{model}} = 1024$, $d_{\text{hidden}} = 4096$, $n_{\text{heads}} = 16$, $n_{\text{layers}} = 6$, and $p_{\text{dropout}} = 0.1$. We adopt the same optimization strategy as Zh/En→Ja except for a larger batch size of 524,288. We train all models for 8 epochs and average the model parameters over the last 5 epochs (see the Appendix for more details). During inference, we use beam search with a beam size of 5 and length penalty of 1.0.


<table>
<thead>
<tr>
<th>Train Size</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fr-En</td>
<td>10.00M</td>
</tr>
<tr>
<td>Cs-En</td>
<td>10.00M</td>
</tr>
<tr>
<td>De-En</td>
<td>4.60M</td>
</tr>
<tr>
<td>Fi-En</td>
<td>4.80M</td>
</tr>
<tr>
<td>Lv-En</td>
<td>1.40M</td>
</tr>
<tr>
<td>Et-En</td>
<td>0.70M</td>
</tr>
<tr>
<td>Ro-En</td>
<td>0.50M</td>
</tr>
<tr>
<td>Hi-En</td>
<td>0.26M</td>
</tr>
<tr>
<td>Tr-En</td>
<td>0.18M</td>
</tr>
<tr>
<td>Gu-En</td>
<td>0.08M</td>
</tr>
</tbody>
</table>

Chinese and Japanese sentences using Jieba\(^4\) and MeCab\(^5\) respectively. We remove duplicated sentence pairs from the training corpora, filter them using langid\(^6\), and filter out sentence pairs whose length ratio exceeds 2.0 using clean-corpus-n.perl\(^7\). We apply byte-pair encoding (Sennrich et al., 2016b) to each language separately with 16K merging operations. Table 1 shows the number of sentence pairs after preprocessing.

\(^4\)https://github.com/fxsjy/jieba
\(^5\)https://taku910.github.io/mecab
\(^6\)https://github.com/saffsd/langid.py
\(^7\)https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl
\(^8\)https://github.com/pytorch/fairseq
We compare our method against the following baseline: MeCab tokenizer before computing BLEU. Multilingual baseline: As shown in Table 3, the bilingual baseline, as it is prone to translation errors in the pivot sentence. Multilingual + pseudo fails to bring further improvements over Multilingual baseline in either direction: it improves BLEU by 0.5 on En→Ja query test set but degrades performance on science and news test sets.

By contrast, our single-encoder bi-source model using single-source inference significantly outperforms Multilingual baseline by 1.2 BLEU and Multilingual + pseudo by 0.7 BLEU on En→Ja query test set, with par performance on other test sets. The multi-encoder variant achieves competitive performance to the single-encoder model on Zh→Ja but obtains significantly lower BLEU on En→Ja. Using bi-source inference with our single-encoder model further improves BLEU by 0.3–0.4 over single-source inference. It significantly outperforms Multilingual baseline by 0.8–1.8 BLEU on En→Ja query and news (out-of-domain) test sets, while achieving on par performance on Zh→Ja and En→Ja science (in-domain) test set. This is probably because English and Japanese are more distant, thus adding a high-quality synthetic Chinese source sentence helps translate the domain-specific English words and phrases that are infrequent in the training data.

To better understand the improvements in BLEU, we conduct the following analysis:

Our model improves accuracy on low-frequency words. We compute the target word F1 binned by frequency in the training data (Neubig et al., 2019) on the three out-of-domain test sets. As shown in Figure 3, on En→Ja where our model obtains the largest BLEU improvements, the largest improvements over the baseline models are on low-frequency words – in the news domain, the largest improvements are on words with frequency between 10 and 50, while in the query domain, it improves more on words with frequency between 50 and 100. It also improves F1 on rare words with frequency below 10, but not as much as for words with frequency above 10. In addition, bi-source
target position, the model simultaneously attends to the source and auxiliary tokens that are semantically correlated. For example, when predicting the word “プロジェクト” (project), the model attends both the English word “project” and the Chinese word “を項目” (project).

3.6 En→X Results

3.6.1 Single-Source Inference

Table 4 shows that the Multilingual baseline outperforms the Bilingual baseline by 2.0 BLEU on average. Consistent with Zh/En→Ja results, Multilingual + pivot underperforms the Multilingual baseline by 10.8 BLEU on average. Our bi-source model with single-source inference further improves over both Multilingual baseline and Multilingual + pseudo on all language pairs. For our model, we only report the BLEU scores for the single-encoder bi-source model, as it yields higher BLEU than the multi-encoder model on Zh/En→Ja translation task.

On the high-resource languages (Fr, Cs, De, Fi, and Lv), simply adding pseudo training data degrades BLEU, while our model improves over the Multilingual baseline by 0.7–1.3 BLEU. On the low-resource languages (Et, Ro, Hi, Tr, and Gu), Multilingual + pseudo outperforms the Multilingual baseline by 0.7 BLEU on average, while our bi-source model further improves over Multilingual + pseudo by 0.9–1.4 BLEU. On average, our model outperforms both Multilingual baseline and Multilingual + pseudo by 1.4 BLEU. We will show in Section 3.6.2 that our model trained on the bi-source model further improves over single-source inference on low-frequency words.

On Zh→Ja news set, the largest gain is on medium-frequency words (Figure in the Appendix).
source objective (Eq. 4) learns more aligned representations across languages, which explains its superiority over the multilingual baselines even with single-source inference.

3.6.2 Bi-Source Inference

As shown in Table 4, adding an auxiliary source sentence improves BLEU over single-source inference on most target languages except French, Czech, and Romanian.\(^{11}\) It achieves an average improvement of +0.5 BLEU over single-source inference and outperforms the multilingual baselines by +1.9 BLEU on average and up to +4.0 BLEU on low-resource languages like Gujarati.

Table 4: BLEU scores of the baseline models and our model with single-source and bi-source inference on En→X translation task. We boldface the top scores. For bi-source inference, we report the average BLEU over the choices of the auxiliary language (excluding the target language).

<table>
<thead>
<tr>
<th></th>
<th>Fr</th>
<th>Cs</th>
<th>De</th>
<th>Fi</th>
<th>Lv</th>
<th>Et</th>
<th>Ro</th>
<th>Hi</th>
<th>Tr</th>
<th>Gu</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilingual baseline</td>
<td>31.8</td>
<td>25.8</td>
<td>33.8</td>
<td>20.6</td>
<td>22.3</td>
<td>13.9</td>
<td>25.2</td>
<td>11.2</td>
<td>12.5</td>
<td>7.8</td>
<td>20.5</td>
</tr>
<tr>
<td>Multilingual baseline</td>
<td>31.1</td>
<td>25.3</td>
<td>33.9</td>
<td>21.4</td>
<td>24.7</td>
<td>19.1</td>
<td>28.3</td>
<td>11.3</td>
<td>17.1</td>
<td>12.6</td>
<td>22.5</td>
</tr>
<tr>
<td>Multilingual + pseudo</td>
<td>30.8</td>
<td>24.7</td>
<td>33.0</td>
<td>20.7</td>
<td>24.3</td>
<td>19.4</td>
<td>28.3</td>
<td>13.0</td>
<td>17.9</td>
<td>13.4</td>
<td>22.5</td>
</tr>
<tr>
<td>Ours (single-source)</td>
<td>31.8</td>
<td>26.5</td>
<td>34.7</td>
<td>22.1</td>
<td>26.0</td>
<td>20.4</td>
<td>29.5</td>
<td>14.2</td>
<td>18.8</td>
<td>14.8</td>
<td>23.9</td>
</tr>
<tr>
<td>Ours (bi-source)</td>
<td>31.6</td>
<td>26.5</td>
<td>35.3</td>
<td>22.3</td>
<td>26.5</td>
<td>21.1</td>
<td>29.1</td>
<td>15.3</td>
<td>19.2</td>
<td>17.4</td>
<td>24.4</td>
</tr>
</tbody>
</table>

Table 5: BLEU scores of our model with single-source and bi-source inference on En→X translation task, and BLEU degradation when 5% and 10% of the source words are masked randomly at inference time.

<table>
<thead>
<tr>
<th></th>
<th>Fr</th>
<th>Cs</th>
<th>De</th>
<th>Fi</th>
<th>Lv</th>
<th>Et</th>
<th>Ro</th>
<th>Hi</th>
<th>Tr</th>
<th>Gu</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our single-source w/o mask</td>
<td>31.8</td>
<td>26.5</td>
<td>34.7</td>
<td>22.1</td>
<td>26.0</td>
<td>20.4</td>
<td>29.5</td>
<td>14.2</td>
<td>18.8</td>
<td>14.8</td>
<td>23.9</td>
</tr>
<tr>
<td>mask 5% source</td>
<td>-5.5</td>
<td>-5.6</td>
<td>-7.5</td>
<td>-5.9</td>
<td>-5.0</td>
<td>-4.8</td>
<td>-5.8</td>
<td>-3.0</td>
<td>-5.7</td>
<td>-1.8</td>
<td>-5.1</td>
</tr>
<tr>
<td>mask 10% source</td>
<td>-10.8</td>
<td>-11.5</td>
<td>-13.7</td>
<td>-11.3</td>
<td>-10.7</td>
<td>-9.6</td>
<td>-10.2</td>
<td>-7.0</td>
<td>-9.5</td>
<td>-4.3</td>
<td>-9.9</td>
</tr>
<tr>
<td>Our bi-source w/o mask</td>
<td>31.6</td>
<td>26.5</td>
<td>35.3</td>
<td>22.3</td>
<td>26.5</td>
<td>21.1</td>
<td>29.1</td>
<td>15.3</td>
<td>19.2</td>
<td>17.4</td>
<td>24.4</td>
</tr>
<tr>
<td>mask 5% source</td>
<td>-5.1</td>
<td>-4.6</td>
<td>-7.4</td>
<td>-5.4</td>
<td>-4.3</td>
<td>-4.7</td>
<td>-5.4</td>
<td>-2.7</td>
<td>-4.9</td>
<td>-2.3</td>
<td>-4.7</td>
</tr>
<tr>
<td>mask 10% source</td>
<td>-9.9</td>
<td>-9.2</td>
<td>-13.4</td>
<td>-9.7</td>
<td>-9.9</td>
<td>-8.9</td>
<td>-9.6</td>
<td>-6.0</td>
<td>-8.6</td>
<td>-5.6</td>
<td>-9.0</td>
</tr>
</tbody>
</table>

\(^{11}\)We use the same model in single-source inference mode to generate the auxiliary sentences.

Typological Analysis To better understand which target language benefits the most from bi-source inference, we compute the Spearman’s correlation between the average BLEU improvement on each target language and various types of features including 1) the training data size for each language pair, and 2) the linguistic distances between the source (English) and the target languages measured by the geographic distance, genetic distance based on the world language family tree, syntactic distance, and phonological distance from URIEL Typological Database (Littell et al., 2017).
Table 6: Average coverage scores of our model with single-source and bi-source inference and the multilingual baseline on MuCoW.

<table>
<thead>
<tr>
<th>Model</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>En→Cs</strong></td>
<td></td>
</tr>
<tr>
<td>Multilingual baseline</td>
<td>56.05</td>
</tr>
<tr>
<td>Ours (single-source)</td>
<td>56.27</td>
</tr>
<tr>
<td>Ours (bi-source)</td>
<td><strong>56.96</strong></td>
</tr>
<tr>
<td><strong>En→De</strong></td>
<td></td>
</tr>
<tr>
<td>Multilingual baseline</td>
<td>57.21</td>
</tr>
<tr>
<td>Ours (single-source)</td>
<td>60.44</td>
</tr>
<tr>
<td>Ours (bi-source)</td>
<td><strong>60.61</strong></td>
</tr>
</tbody>
</table>

Results show that the geographic distance correlates the best with the BLEU improvement with a correlation score of 0.74, which suggests that more distant language pairs benefit more from the auxiliary source sentences. In addition, the BLEU improvement correlates negatively with the training data size with a correlation score of -0.57, which suggests that lower-resource language pairs obtain a larger gain from bi-source inference. The genetic, syntactic, and phonological distances do not correlate well with the BLEU improvement.\(^{12}\)

Word Sense Disambiguation To test if bi-source inference helps disambiguate word senses, we compare our En→X model with single-source and bi-source inference with the *Multilingual baseline* on the MuCoW test suite (Raganato et al., 2019), a word sense disambiguation test suite. Table 6 shows that our model with bi-source inference achieves higher coverage scores over its counterpart with single-source inference and *Multilingual baseline* on both En→Cs and En→De. This confirms our hypothesis that adding an auxiliary language input during inference helps disambiguate word senses.

4 Related Work

Since the recent success of the end-to-end NMT models (Sutskever et al., 2014; Bahdanau et al., 2015), multilingual NMT has become a promising research direction. Dong et al. (2015) propose to perform one-to-many translation using a dedicated decoder for each target language. Firat et al. (2016a) further extend it to support many-to-many translation using language-specific encoders and decoders with a shared attention module. Ha et al. (2016) and Johnson et al. (2017) show that training a shared encoder-decoder model for many-to-many translation allows translation between unseen language pairs. More advanced techniques to further improve the translation quality include optimizing the parameter sharing strategies (Gu et al., 2018; Sachan and Neubig, 2018) and multi-stage fine-tuning to better improve low-resource translation (Dabre et al., 2019). Although we only focus on improving the overall translation quality of a shared multilingual NMT model in this paper, our approach can also be combined with the aforementioned techniques to build better language-specific NMT models via fine-tuning, which we will explore in future work.

Orthogonal to these techniques, multi-source translation (Och and Ney, 2001; Zoph and Knight, 2016; Garmash and Monz, 2016) has been shown to improve translation quality by exploiting the source sentences manually translated into multiple languages. Most studies assume access to multi-source inputs during both training and inference. Choi et al. (2018) and Nishimura et al. (2018) introduce data augmentation methods to fill in the missing source in the training data. Firat et al. (2016b) explore translating the source into a pivot language and feeding both the original source and pivot sentences to a multilingual model to improve zero-resource translation. However, the pivot sentence is added only at inference time, thus the approach is better suited to the zero-resource setting. More recently, Tuitelbaum et al. (2019) shows that translating the source word to auxiliary languages improves word translation.

Our work is also related to multi-task learning for machine translation. Tu et al. (2017) propose multi-task learning with an auxiliary reconstruction objective that reconstructs the source sentence from decoder hidden states. Niu et al. (2019) further show that adding a reconstruction objective by back-translating the target sentences to the source helps low-resource translation. Zhou et al. (2019) propose multi-task training with a denoising objective to improve the robustness of NMT models. Wang et al. (2020) show that multi-task learning with two additional denoising tasks on the monolingual data can effectively improve translation quality. Our training strategy can also be viewed as multi-task learning as we train our multilingual model on single-source and bi-source inputs jointly.

\(^{12}\)We assume that the correlation is weak if the absolute correlation score is below 0.4.
5 Conclusion

We introduced a novel bi-source multilingual translation model that exploits an additional source input from an auxiliary language to improve translation quality. Our model can flexibly perform single-source and bi-source inference, in which it takes both the original source and a synthetic source sentence from an auxiliary language as inputs. Experiments show that our method is simple yet effective—it improves the translation quality of multilingual models substantially, with the largest improvements on low-resource or distant language pairs. Further analysis indicates that adding an auxiliary language input during inference helps the model disambiguate source words. This work also sheds new light on multilingual NMT training, as our multi-source training strategy brings substantial improvements over the multilingual baseline without adding any auxiliary inputs at inference time.

Acknowledgement

We thank the anonymous reviewers, Eleftheria Briakou, Naomi Feldman, Nika Jurov, Yusen Lin, the CLIP lab, and Department of Linguistics at UMD for their valuable feedback.

References


Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. 2016. Ensem-

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Ensem-


A Model and Training Details

Table 7 shows the total number of parameters for each model. For the Zh/En→Ja task, training each model takes 36 hours on 4 NVIDIA Tesla P40 GPUs for hours. For the En→X task, training each model takes around 72 hours on 8 V100 GPUs.

<table>
<thead>
<tr>
<th>Model Size (M)</th>
<th>Zh/En→Ja</th>
<th>En→X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bilingual baseline</td>
<td>9.6 / 241.9</td>
</tr>
<tr>
<td></td>
<td>Multilingual baseline</td>
<td>241.9</td>
</tr>
<tr>
<td></td>
<td>Multilingual + pseudo</td>
<td>241.9</td>
</tr>
<tr>
<td></td>
<td>Ours (multi-enc)</td>
<td>241.9</td>
</tr>
<tr>
<td></td>
<td>Ours (single-enc)</td>
<td>241.9</td>
</tr>
</tbody>
</table>

Table 7: Model sizes (M) for Zh/En→Ja and En→X tasks. For the bilingual baseline on En→X, we report the model sizes for the low-resource (Tr, Hi, and Gu) and high-resource languages (Fr, Cs, De, Fi, Lv, Et, Ro), separately.

B Word F1 versus Frequency on Zh→Ja

Figure 6: Target word F1 score binned by word frequency in training data on Zh→Ja.