Improving Multilingual Neural Machine Translation with Auxiliary Source Languages

Weijia Xu[*1] **Yuwei Yin**[*2] **Shuming Ma**[3] **Dongdong Zhang**[3] **Haoyang Huang**[3]
[1]Department of Computer Science, University of Maryland
[2]State Key Lab of Software Development Environment, Beihang University
[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.

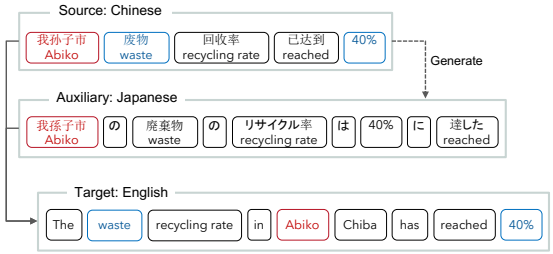


Figure 1: An example of translating a Chinese sentence into English by using Japanese as the auxiliary language. Adding a synthetic source from Japanese helps to translate the red word "我孙子市" (Abiko, a city in Japan), which is often incorrectly translated into "my grandson city" by standard Chinese-English MT models, while other words can be translated more accurately from the Chinese 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

[*] Contribution during internship at Microsoft Research Asia.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3029–3041
November 7–11, 2021.

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 $\boldsymbol{y}^{l_t}$ in language l_t given the original source sentences $\boldsymbol{x}^{l_s}$ from language l_s and a synthetic source sentence $\tilde{\boldsymbol{x}}^{l_a}$ translated from $\boldsymbol{x}^{l_s}$ into an auxiliary language l_a ($l_a \neq l_s$, $l_a \neq l_t$) by an MT model:

$$p(\boldsymbol{y}^{l_t} \mid \boldsymbol{x}^{l_s}, \tilde{\boldsymbol{x}}^{l_a}) = p(\boldsymbol{y}^{l_t} \mid f(\boldsymbol{x}^{l_s}, \tilde{\boldsymbol{x}}^{l_a}; \Theta_{enc}); \Theta_{dec}) \quad (1)$$

where Θ_{enc} and Θ_{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 $f(\boldsymbol{x}^{l_s}; \Theta_{enc}) = \boldsymbol{H}^s$ and $f(\tilde{\boldsymbol{x}}^{l_a}; \Theta_{enc}) = \boldsymbol{H}^a$. Then, the decoder can attend to $\boldsymbol{H}^s$ and $\boldsymbol{H}^a$ separately and apply a gating mechanism to obtain the fusion vector $\boldsymbol{h}_i$:

$$\begin{aligned} \boldsymbol{h}_i^s &= Attn(\boldsymbol{H}^s, \boldsymbol{h}_i^{tgt}) \\ \boldsymbol{h}_i^a &= Attn(\boldsymbol{H}^a, \boldsymbol{h}_i^{tgt}) \\ \boldsymbol{g}_i &= \sigma(\boldsymbol{W}_g[\boldsymbol{h}_i^s; \boldsymbol{h}_i^a] + \boldsymbol{b}_g) \\ \boldsymbol{h}_i &= \boldsymbol{g}_i \odot \boldsymbol{h}_i^s + (1 - \boldsymbol{g}_i) \odot \boldsymbol{h}_i^a \end{aligned} \quad (2)$$

where $\boldsymbol{h}_i^{tgt}$ represents the hidden state of the i-th target token, $\boldsymbol{W_g}$ and $\boldsymbol{b}_g$ are model parameters, and σ 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[1] and a stack of self-attention and position-wise feed-forward layers to produce a sequence of hidden representations $f([\boldsymbol{x}^{l_s}; \tilde{\boldsymbol{x}}^{l_a}]; \Theta_{enc}) = \boldsymbol{H}$. Then, we apply the encoder-decoder attention to the full sequence of encoder representations $\boldsymbol{H}$:

$$\boldsymbol{h}_i = Attn(\boldsymbol{H}, \boldsymbol{h}_i^{tgt}) \quad (3)$$

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 $\mathcal{D} = \bigcup_{l_s \in \mathcal{S}, l_a \in \mathcal{A}, l_t \in \mathcal{T}} \mathcal{D}^{l_s \times l_a \times l_t}$, 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 $\mathcal{D}^{l_s \times l_a \times l_t} = \{(\boldsymbol{x}^{l_s}, \tilde{\boldsymbol{x}}^{l_a}, \boldsymbol{y}^{l_t})\}$ is a bi-source translation dataset which can be formed by data augmentation via MT. The objective is to

[1] The position embeddings of the source sentences are reset to facilitate alignment between two sentences.

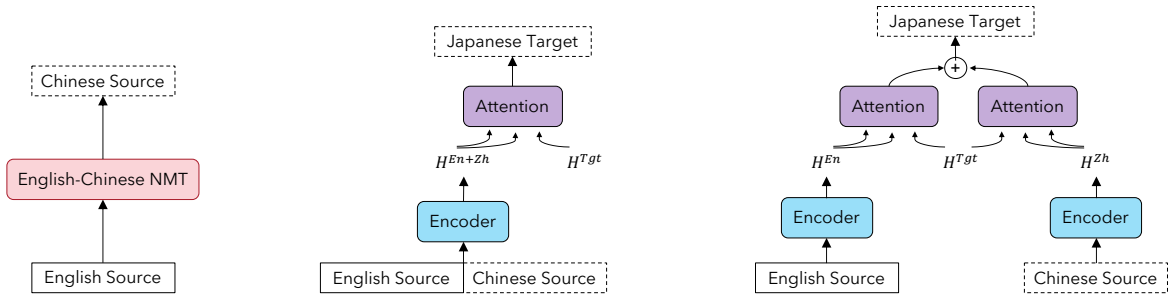


Figure 2: An overview of the generation process of the auxiliary source sentence (a), the single-encoder (b) and multi-encoder (c) approaches for integrating the auxiliary source sentence in the translation model. In the multi-encoder approach, we share the parameters of the two encoders to learn representations in a shared space.

maximize the log-likelihood of the target sentences given the original and auxiliary source sentences:

$$\mathcal{L} = \sum_{(\boldsymbol{x}^{l_s}, \tilde{\boldsymbol{x}}^{l_a}, \boldsymbol{y}^{l_t}) \in \mathcal{D}} \Big[p_{\text{mask}} \log p(\boldsymbol{y}^{l_t} \mid \boldsymbol{x}^{l_s}) + (1 - p_{\text{mask}}) \log p(\boldsymbol{y}^{l_t} \mid \boldsymbol{x}^{l_s}, \tilde{\boldsymbol{x}}^{l_a}) \Big] \quad (4)$$

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 $\{(\boldsymbol{x}^{l_s}, \tilde{\boldsymbol{x}}^{l_a}, \boldsymbol{y}^{l_t})\}$ from $\mathcal{D}^{l_s \times l_a \times l_t}$ and maximize the log probability of the target sentence $\boldsymbol{y}^{l_t}$ given source sentence $\boldsymbol{x}^{l_s}$ and auxiliary sentence $\tilde{\boldsymbol{x}}^{l_a}$. To enable more flexible decoding and to improve model robustness, we randomly mask out the auxiliary sentences with probability p_{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 $\mathcal{M}_{\mathcal{S} \to \mathcal{A}}$ to translate between source and auxiliary languages. Next, we extend each parallel dataset $\{(\boldsymbol{x}^{l_s}, \boldsymbol{y}^{l_t})\}$ to pseudo bi-source datasets $\{(\boldsymbol{x}^{l_s}, \tilde{\boldsymbol{x}}^{l_a}, \boldsymbol{y}^{l_t})\}$ by translating $\boldsymbol{x}^{l_s}$ into auxiliary languages l_a using $\mathcal{M}_{\mathcal{S} \to \mathcal{A}}$. Finally, we combine all pseudo bi-source datasets into the training data $\mathcal{D}$ 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 $\mathcal{M}_{\mathcal{S} \to \mathcal{A}}$ and then use our bi-source model to generate the target translation given the original and auxiliary source sentences.

[2]We set $p_{\text{mask}} = 0.5$ in all our experiments.

		Domain	Prov.	#Sent
Zh-Ja	train	Science	ASPEC	0.66M
	dev	Science	ASPEC	2090
	test	Science	ASPEC	2107
		News	Internal	1000
En-Ja	train	Science	ASPEC	2.63M
	dev	Science	ASPEC	1790
	test	Science	ASPEC	1812
		Query	Internal	4999
		News	WMT20	993
Zh-En	train	News	WMT18	18.7M
	dev	News	WMT18	2001

Table 1: Domain, Provenance (*Prov.*), and the number of sentence pairs (*#Sent*) in the training, development, and test data for Zh-Ja, En-Ja, and Zh-En.

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-

	Train Size	Test
Fr-En	10.00M	newstest13
Cs-En	10.00M	newstest16
De-En	4.60M	newstest16
Fi-En	4.80M	newstest16
Lv-En	1.40M	newsdev17
Et-En	0.70M	newsdev18
Ro-En	0.50M	newsdev16
Hi-En	0.26M	newsdev14
Tr-En	0.18M	newstest16
Gu-En	0.08M	newsdev19

Table 2: Number of sentence pairs in the training data and the test set for each language pair.

vian (Lv), Estonian (Et), Romanian (Ro), Hindi (Hi), Turkish (Tr), and Gujarati (Gu).

Zh/En→Ja We set the source and auxiliary language sets $\mathcal{S} = \mathcal{A} = \{\text{Zh}, \text{En}\}$, and the target language set $\mathcal{T} = \{\text{Ja}\}$. 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 set the source language set $\mathcal{S} = \{\text{En}\}$, the auxiliary and target language sets $\mathcal{A} = \mathcal{T} = \{\text{Fr}, \text{Cs}, \text{De}, \text{Fi}, \text{Lv}, \text{Et}, \text{Ro}, \text{Hi}, \text{Tr}, \text{Gu}\}$. The training data are from the WMT corpus (Bojar et al., 2013, 2014, 2016, 2017, 2018; Barrault et al., 2019).[3] We use all the available parallel data except for the WikiTitles released by WMT19. For French and Czech, we randomly sample 10M sentence pairs from the full data.

3.2 Preprocessing

Zh/En→Ja We tokenize the English sentences using Moses (Koehn et al., 2007) and segment 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.

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.[8]

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.

[3] Data can be downloaded from http://www.statmt.org/wmt13/translation-task.html, http://statmt.org/wmt14/translation-task.html, http://www.statmt.org/wmt16/translation-task.html, http://www.statmt.org/wmt17/translation-task.html, http://www.statmt.org/wmt18/translation-task.html, and http://www.statmt.org/wmt19/translation-task.html

[4] https://github.com/fxsjy/jieba

[5] http://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

	Inference	Zh→Ja Science	Zh→Ja News	Zh→Ja Avg	En→Ja Science	En→Ja Query	En→Ja News	En→Ja Avg
Bilingual baseline	–	46.6	18.7	32.6	**43.1**	12.1	9.2	21.5
Multilingual baseline	–	47.6	20.6	34.1	42.7	13.3	9.8	21.9
Multilingual + pseudo	–	47.5	20.1	33.8	42.3	13.8	9.2	21.8
Multilingual + pivot	–	20.5	7.7	14.1	19.4	10.8	6.7	12.3
Ours (multi-enc)	single	47.6	20.5	34.1	42.3	13.0	8.4	21.2
Ours (single-enc)	single	48.0	20.1	34.1	42.6	14.5	10.1	22.4
Ours (multi-enc)	bi-source	47.8	**20.9**	**34.5**	42.8	14.5	**10.6**	22.6
Ours (single-enc)	bi-source	**48.1**	20.8	34.4	42.7	**15.1**	**10.6**	**22.8**

Table 3: BLEU scores on Zh/En→Ja translation task. We compare our models in the single-source and bi-source inference modes. We boldface the highest scores and underline their ties based on paired bootstrap with $p < 0.05$ (Clark et al., 2011). Our model with bi-source inference significantly outperforms both the *Multilingual baseline* and *Multilingual + pseudo* on En→Ja, and achieves on par performance on Zh→Ja.

3.4 Baselines and Evaluation

We compare our method against the following baselines: 1) **Bilingual baseline:** NMT model trained on each language pair separately. 2) **Multilingual baseline:** multilingual NMT model trained on Zh-Ja and En-Ja data for Zh/En→Ja, and all English-centric data for En→X. 3) **Multilingual + pseudo:** multilingual NMT model trained on the concatenation of the original parallel data $\{(\boldsymbol{x}^{l_s}, \boldsymbol{y}^{l_t})\}$ and pseudo data $\{(\tilde{\boldsymbol{x}}^{l_a}, \boldsymbol{y}^{l_t})\}$. 4) **Multilingual + pivot:** multilingual NMT model with pivot decoding (by first translating the source to the auxiliary language and then translating from the auxiliary to the target language). For all multilingual models, we add the target language tag and temperature-based sampling (Aharoni et al., 2019) with temperature $\tau = 5$. We evaluate translation quality using sacreBLEU (Post, 2018).[9] For Japanese, we use *MeCab* tokenizer before computing BLEU.

3.5 Zh/En→Ja Results

As shown in Table 3, *Multilingual baseline* outperforms *Bilingual baseline* by 0.4–1.5 BLEU on average, while *Multilingual + pivot* underperforms *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.

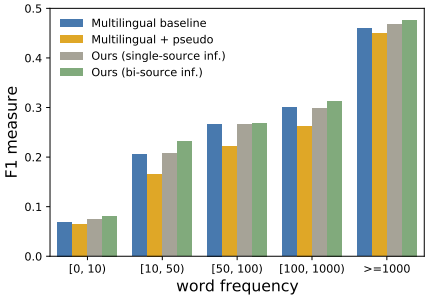
By contrast, our single-encoder bi-source model using single-source inference significantly outperforms *Multilingual baseline* by 1.2 BLEU and *Multilingual + pseudo* by 0.7 BLEU on En→Ja query test set, with on par performance on other test sets.[10] The multi-encoder variant achieves competitive performance 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:

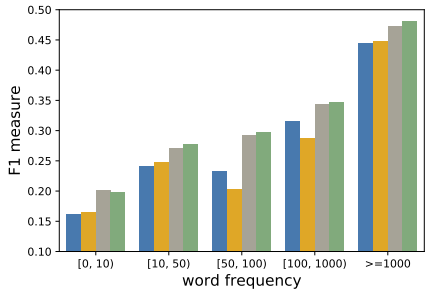
Our model improves accuracy on low-frequency words. We compute the target word F1 binned by frequency in the training data (Neubig et al., 2019) on the three out-of-domain test sets. As shown in Figure 3, on En→Ja where our model obtains the largest BLEU improvements, the largest improvements over the 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

[9] Version: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.3

[10] All mentions of significance are based on the paired bootstrap test (Clark et al., 2011) with $p < 0.05$.



(a) En→Ja News



(b) En→Ja Query

Figure 3: Target word F1 score binned by word frequency in training data on En→Ja. Our model improves the most over multilingual baselines on low-frequency words.

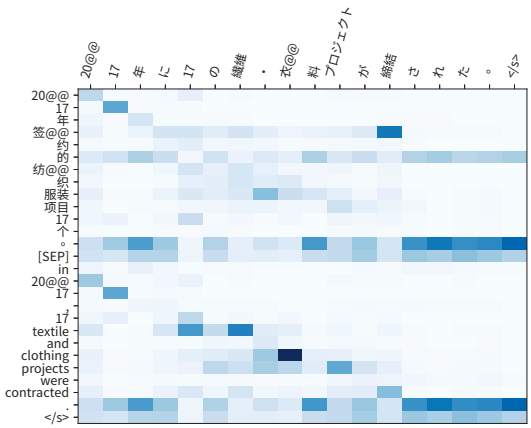


Figure 4: Visualizing the encoder-decoder attention weights (averaged over all attention heads) of the single-encoder bi-source model for an example from the Zh→Ja news test set. Our model learns the alignments between words in the source and auxiliary sentences.

inference improves over single-source inference more on low-frequency words on En→Ja news set. On Zh→Ja news set, the largest gain is on medium-frequency words (Figure in the Appendix).

Our model learns the alignments between source and auxiliary tokens. We examine the encoder-decoder attention weights of our single-encoder bi-source model. Figure 4 shows a typical example from the Zh→Ja news test set. At each

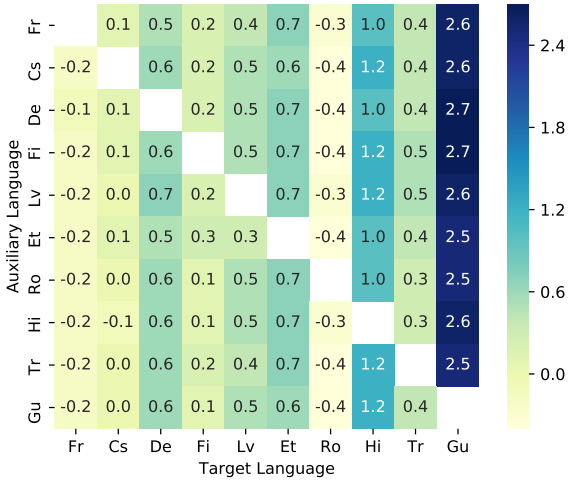


Figure 5: BLEU improvements of bi-source inference using different auxiliary languages over single-source inference on En→X translation task.

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 "项目" (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-

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

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).

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

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.

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.

Figure 5 shows the BLEU improvements from adding different auxiliary languages over single-source inference. The choice of the auxiliary language has little impact on the BLEU improvement. To explain this phenomena, we conduct the following analysis to verify that the performance boosts are due to the additional source information provided by the auxiliary sentence. We compare single-source and bi-source inference on synthetic noisy test sets: we randomly mask τ% of the source words in each test set ($\tau \in \{5, 10\}$). As show in Table 5, when using single-source inference, BLEU drops by -5.1 and -9.9 after masking 5% and 10% of the source words, respectively. With the help of the auxiliary language, the drop in BLEU becomes smaller: the drop is reduced by 0.4 and 0.9 when 5% and 10% of the source words are masked, respectively. These results indicate that **our model can effectively leverage the complementary information provided by the auxiliary sentence** which remedies the missing source information. Furthermore, results suggest that the cross-lingual representations in our model are well-aligned which enables it to combine the information from both the source and auxiliary sentences. This also explains why the choice of the auxiliary language has little impact on BLEU – as the representations of the auxiliary sentences from different languages are close in the hidden space, they could complement the source context similarly.

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).

[11] We use the same model in single-source inference mode to generate the auxiliary sentences.

	Model	Coverage
En→Cs	Multilingual baseline	56.05
	Ours (single-source)	56.27
	Ours (bi-source)	**56.96**
En→De	Multilingual baseline	57.21
	Ours (single-source)	60.44
	Ours (bi-source)	**60.61**

Table 6: Average coverage scores of our model with single-source and bi-source inference and the multilingual baseline on MuCoW.

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, Taitelbaum 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

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3th International Conference on Learning Representations*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, 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: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation*, pages 272–307. Association for Computational Linguistics.

Gyu-Hyeon Choi, Jong-Hun Shin, and Young-Kil Kim. 2018. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA. Association for Computational Linguistics.

Raj Dabre, Fabien Cromierès, and Sadao Kurohashi. 2017. Enabling multi-source neural machine translation by concatenating source sentences in multiple languages. *CoRR*, abs/1702.06135.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416, Hong Kong, China. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 344–354. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3th International Conference on Learning Representations*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Asso-*

ciation for Computational Linguistics (Demonstrations), pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with missing data. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 92–99, Melbourne, Australia. Association for Computational Linguistics.

Xing Niu, Weijia Xu, and Marine Carpuat. 2019. Bidirectional differentiable input reconstruction for low-resource neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 442–448, Minneapolis, Minnesota. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *MT summit*, pages 253–258.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The mucow test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 470–480. Association for Computational Linguistics.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112.

Hagai Taitelbaum, Gal Chechik, and Jacob Goldberger. 2019. Multilingual word translation using auxiliary languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1330–1335, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3097–3103. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, Long Beach, CA, USA. Curran Associates, Inc.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving robustness of neural machine translation with multi-task learning. In *Proceedings of the Fourth Conference on Machine Translation (Volume*

2: Shared Task Papers, Day 1), pages 565–571, Florence, Italy. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

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.

	Model Size (M)
Zh/En→Ja	
Bilingual baseline	60.9
Multilingual baseline	69.1
Multilingual + pseudo	69.1
Ours (multi-enc)	69.1
Ours (single-enc)	69.1
En→X	
Bilingual baseline	9.6 / 241.9
Multilingual baseline	241.9
Multilingual + pseudo	241.9
Ours (single-enc)	241.9

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

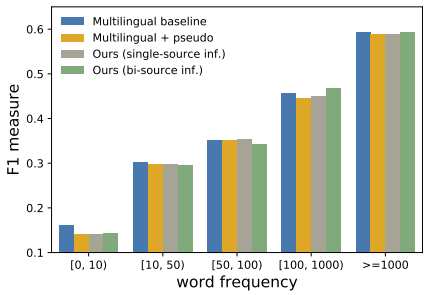


Figure 6: Target word F1 score binned by word frequency in training data on Zh→Ja.