Improving Multilingual Neural Machine Translation For Low-Resource Languages: French, English - Vietnamese

Thi-Vinh Ngo
Thai Nguyen University
ntvinh@ictu.edu.vn

Phuong-Thai Nguyen
Vietnam National University
npthai@vnu.edu

Thanh-Le Ha
Karlsruhe Institute of Technology
thanh-le.ha@kit.edu

Khac-Quy Dinh
Vietnam National University
moduledk@gmail.com

Le-Minh Nguyen
JAIST, Japan
nguyenml@jaist.ac.jp

Abstract

Prior works have demonstrated that a low-resource language pair can benefit from multilingual machine translation (MT) systems, which rely on many language pairs' joint training. This paper proposes two simple strategies to address the rare word issue in multilingual MT systems for two low-resource language pairs: French-Vietnamese and English-Vietnamese. The first strategy is about dynamical learning word similarity of tokens in the shared space among source languages while another one attempts to augment the translation ability of rare words through updating their embeddings during the training. Besides, we leverage monolingual data for multilingual MT systems to increase the amount of synthetic parallel corpora while dealing with the data sparsity problem. We have shown significant improvements of up to +1.62 and +2.54 BLEU points over the bilingual baseline systems for both language pairs and released our datasets for the research community.

1 Introduction

Neural Machine Translation (NMT) (Bahdanau et al., 2015) has achieved state of the art in various MT systems, including rich and low resource language pairs (Edunov et al., 2018; Gu et al., 2019; Ngo et al., 2019). However, the quality of low-resource MT is quite unpretentious due to the lack of parallel data while it has achieved better results on systems of the available resource. Therefore, low-resource MT is one of the essential tasks investigated by many previous works (Ha et al., 2016; Lee et al., 2016; Sennrich and Zhang, 2019).

Recently, some works present MT systems that have achieved remarkable results for low-resource language (Gu et al., 2019; Aharoni et al., 2019). Inspired by these works, we collect data from the TED Talks domain, then attempt to build multilingual MT systems from French, English-Vietnamese. Experiments demonstrate that both language pairs: French-Vietnamese and English-Vietnamese have achieved significant performance when joining the training.

Although multilingual MT can reduce the sparse data in the shared space by using word segmentation, however, rare words still exist, evenly they are increased more if languages have a significant disparity in term vocabulary. Previous works suggested some strategies to reduce rare words such as using translation units at sub-word and character levels or generating a universal representation at the word and sentence levels (Lee et al., 2016; Gu et al., 2019). These help to downgrade the dissimilarity of tokens shared from various languages. However, these works require learning additional parameters in training, thus increasing the size of models.

Our paper presents two methods to augment the translation of rare words in the source space without modifying the architecture and model size of MT systems: (1) exploiting word similarity. This technique has been mentioned by previous works (Luong et al., 2015; Li et al., 2016; Trieu et al., 2016; Ngo et al., 2019). They employ monolingual data or require supervised resources like a bilingual dictionary or WordNet, while we leverage relation from the multilingual space of MT systems. (2) Adding a scalar value to the rare word embedding in order to facilitate its translation in the training process.

Due to the fact that NMT tends to have bias in translating frequent words, so rare words (which have low frequency) often have less opportunity to be considered. Our ideal is inspired by the works of (Nguyen and Chiang, 2017; Ngo et al., 2019; Gu et al., 2019). (Nguyen and Chiang, 2017) and (Ngo et al., 2019) proposed various solutions to urge for translation of rare words, including modification

Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, pages 55–61
Decmber 04, 2020. ©2020 Association for Computational Linguistics

embedding in training. They only experimented with recurrent neural networks (RNNs) while our work uses the state-of-the-art transformer architecture. (Gu et al., 2019) transforms the word embedding of a token into the universal space, and they learn plus parameters while our method does not. We apply our strategies in our fine-tuning processes, and we show substantial improvements of the systems after some epochs only.

Monolingual data are widely used in NMT to augment data for low-resource NMT systems (Sennrich et al., 2015; Zhang and Zong, 2016; Lample et al., 2018; Wu et al., 2019; Siddhant et al., 2020). Back-translation (Sennrich et al., 2015) is known as the most popular technique in exploiting target-side monolingual data to enhance the translation systems while the self-learning method (Zhang and Zong, 2016) focuses on utilizing source-side monolingual data. Otherwise, the dual-learning strategy (Wu et al., 2019) also suggests using both source- and target-side monolingual data to tackle this problem. Our work investigates the self-learning method (Zhang and Zong, 2016) on the low-resource multilingual NMT systems specifically related to Vietnamese. Besides, monolingual data are also leveraged in unsupervised(Lample et al., 2018) or zero-shot translation(Lample et al., 2018).

The main contributions of our work are:

- We first attempt to build a multilingual system for two low-resource language pairs: French-Vietnamese and English-Vietnamese.
- We propose two simple techniques to encourage the translation of rare words in multilingual MT to upgrade the systems.
- We investigate the quality translation of the low-resource multilingual NMT systems when they are reinforced synthetic data.
- We release more datasets extracted from the TED Talks domain for the research purpose: French-Vietnamese and English-Vietnamese.

In section 2, we review the transformer architecture used for our experiments. The brief of multilingual translation is shown in section 3. Section 4 presents our methods to deal with rare words in multilingual translation scenarios. The exploitation of monolingual data for low-resource multilingual MT is discussed in section 5. Our results are described in section 6, and related work is shown in section 7. Finally, the paper ends with conclusions and future work.

2 Transformer-based NMT

Transformer architecture for machine translation is mentioned for the first time by (Vaswani et al., 2017). This is based on the sequence to sequence framework (Sutskever et al., 2014) which includes an encoder to transform information of the source sentence $X = (x_1, x_2, ..., x_n)$ into continuous representation and a decoder to generate the target sentence $Y = (y_1, y_2, ..., y_m)$.

Self-attention is an important mechanism in the transformer architecture. It enables the ability to specify the relevance of a word with the remaining words in the sentence through the equation:

$$\text{Self-Attn}(Q, K, V) = \text{Softmax}(\frac{QK^T}{d})V \quad (1)$$

where *K* (key), *Q* (query), *V* (value) are the representations of the input sentence and *d* is the size of the input. The attention mechanism (Luong et al., 2015a) bridges between the source sentence in the encoder and the target sentence in the decoder. Furthermore, the feed-forward networks are used to normalize the outputs on both encoder and decoder.

The MT system is trained to minimize the maximum likelihood of *K* parallel pairs:

$$\mathcal{L}(\theta) = \frac{1}{K}\sum_{k=1}^{k=K} logp(Y^k|X^k;\theta) \quad (2)$$

3 Multilingual NMT

Multilingual NMT systems can translate between many language pairs, even in the zero-shot issue. Previous works investigate multilingual translation in many fashions: (1) Many to many (Ha et al., 2016; Aharoni et al., 2019): from many sources to many target languages; (2) Many to one (Gu et al., 2019): from many source languages to a target language; (3) One to many (Wang et al., 2018): from one source language to many target languages. In cases (1) and (3), an artificial token is often added to the beginning of the source sentence to specify the predicted target language. Our MT systems are the same as the case (2), so we do not add any artificial token to the texts.

In a multilingual NMT system from many to one with M language pairs and K sentence pairs for each one, the objective function uses maximum

likelihood estimation on the whole parallel pairs $\{X^{(m,k)}, Y^{(m,k)}\}_{k=1..K}^{m=1...M}$ as:

$$\mathcal{L}(\theta) = \frac{1}{K} \sum_{m=1}^{m=M} \sum_{k=1}^{k=K} logp(Y^{(m,k)}|X^{(m,k)}; \theta) \quad (3)$$

where $K = \sum_{m=1}^{m=M} K_m$ is the total number of sentences of the whole corpus.

The vocabulary of the source side is mixed from all source languages: $V = \sum_{m=1}^{m=M} V_m$.

(Gu et al., 2019) has shown that if the languages shared the same alphabet and had many similar words, such system will get many advantages from multilingual MT. In fact, different words from many languages can differ in form, but they may share the same subwords. This significantly reduces the number of rare words in the MT systems. Nevertheless, the rare word issue is still a challenge in NMT. We choose English and French are source languages in our experiment with the hope that they can share many tokens even though we do not have much data of those translation directions.

4 Augmenting Rare Word Translation

4.1 Learning multilingual word similarity

We assume that a rare word or rare token (which has a low frequency in the training data) from one source language may be similar to another word in a shared multilingual space. Similar words can belong to several languages and they can be replaced by the others.

Our method replaces rare tokens with their similar tokens in shared space. The replacements are learned dynamically in the training NMT system. To avoid slowing down the training speed, we only compute similar tokens after each epoch. In the experiments, we attempt to replace rare tokens from French with similar tokens in English and French.

Our method is described as follows:

Firstly, we extract the lists of all tokens from the English - $\{A\}$ corpus, and the most *k* common words from the vocabulary of the source side of the French - $\{B\}$. We set *k=15* thousand words in the experiments.

Secondly, we compute the similarity score between the embedding of a rare token t_i, $\forall t_i \notin \{A \cup B\}$ and each embedding of the tokens t_j, $\forall t_j \in \{A \cup B\}$ as follows:

$$score_i = min(d_j(e_i, e_j) \cdot e^{\cos(e_i, e_j)}) \quad (4)$$

where $j = 1..M$ with M is the number of tokens of $A \cup B$; d is the Euclidean distance between embedding e_i of token t_i and embedding e_j of token t_j.

The last, the token t_i is replaced by its similar tokens. The scores are computed iteratively after each epoch during the training process. It may have more tokens similar to a rare token, so we experimentalize in the case of random selection a token from the similar tokens. To accrete the effectiveness of the method, we use a threshold to neglect similar pairs that have scores close to 0 or too large. In the experiments, we choose the scores in $[2.4, 2.72]$ to warrant similar pairs alike in terms of distance as well as direction.

4.2 Updating source embedding

In this approach, we assume that the embedding e_i of token t_j, $\forall t_i \notin \{A \cup B\}$ is represented by the approximate embedding vector as following:

$$e_i = e_i + d \quad (5)$$

where d is the difference between embedding e_i and the average of the all embeddings e_j of token t_j, $\forall t_j \in \{A \cup B\}$:

$$d = e_i - \frac{\sum_{j=1}^{j=M} e_j}{M} \quad (6)$$

where M is the number of tokens of $\{A \cup B\}$.

These embeddings are then updated during the training. The average of embeddings is only estimated after each epoch to avoid slowing down the training speed. We observe the improvements in both language pairs in the experiments.

5 Exploiting monolingual data for low-resource multilingual NMT

Similar to the idea suggested in (Zhang and Zong, 2016), we leverage monolingual data from the source-side to generate synthetic bilingual data. Instead of using monolingual data from all source languages, we only attempt to exploit monolingual data of English.

Firstly, we train the multilingual NMT system from English, French → Vietnamese based on bilingual data from the TED talks with the approaches mentioned in section 4. The best system is then used to translate English to Vietnamese.

Lastly, the synthetic parallel data are mixed with original bilingual data in the normal training scheme.

6 Experiments

6.1 Datasets

We extracted data from TED Talks domain[1] for two language pairs English-Vietnamese and French-Vietnamese. The details of those datasets are described in Table 1. For the English-Vietnamese, we used standard datasets like `tst2012` and `tst2013` from (Cettolo et al., 2016) as dev and test sets for validation and evaluation. For French-Vietnamese, we separate a subset from collected data for the same purposes.

Datasets	Training	dev	test
English-Vietnamese	231K	1553	1268
French-Vietnamese	203K	1007	1049

Table 1: The bilingual datasets in our experiments

To generate synthetic bilingual data, we sampled 1.2 millions English monolingual sentences from the European Parliament English-French corpus[2]. After inferring from the multilingual MT system, we obtained two sets of pseudo bilingual data: English - Vietnamese, French - Vietnamese.

6.2 Preprocessing

English and French texts were tokenized and true-cased using Moses's scripts, and then they are applied to Sennrich's BPE (Sennrich et al., 2016). 30000 operators are learned to generate BPE codes for both languages.

For Vietnamese texts, we only did tokenization and true-casing using Moses's scripts.

We extracted a list of all tokens in English (A) and another list of the 15K most frequency of tokens in French (B). All lists were then used for the mentioned strategies in section 4.

6.3 Systems and Training

We implement our NMT systems using the framework `NMTGMinor`[3]. The same settings are used for all experiments. The system includes 4 layers for both encoder and decoder, and the embedding size is 512. For the systems that adapted monolingual data, we use 6 layers. Adam optimizer is set with the initial learning rate at 1.0 for baseline and the multilingual systems and 0.5 for the fine-tuned systems. The size of a mini-batch is 128, and the vocabulary size is set to be the top 50K most frequent tokens. Training and development sets of both language pairs are concatenated prior to the training of our multilingual systems.

We modified this framework to apply our ideals proposed in section 4. To speed up the training, we compute the similarity scores and find out similar tokens for rare tokens or the mean of all tokens in $\{A \cup B\}$ after each epoch. We replace rare tokens or update their embeddings in each batch. We do not use these techniques for the decoding process, so the system's performance is not affected.

The baseline and multilingual systems are trained for 70 epochs. Our methods are then used to fine-tune the systems for 15 epochs. We choose the five best models to decode the test sets independently for residual systems despite the baseline systems. The beam size is 10, and we try different values of *alpha*: $0.2, 0.4, 0.8, 1.0$. Other settings are the default settings of `NMTGMinor`.

6.4 Results

We evaluate the quality of systems on two translation tasks: French to Vietnamese and English to Vietnamese, using on different approaches mentioned in previous sections. The `multi-BLEU` from Moses's scripts[4] is used. The results have shown in the Table 2.

(1) Bilingual baseline systems. We train the systems based on separate bilingual data of each language pair for 70 epochs. The best model is used to decode the test data for comparison purposes in our experiments.

(2) Multilingual systems. We concatenate training and development sets in order to construct the new sets: French, English $\rightarrow$ Vietnamese, and then train the system using those data for the same number of epochs as for the baseline systems. We observe an improvement of +1.05 BLEU points on English $\rightarrow$ Vietnamese translation task and another one of +1.19 BLEU points on French $\rightarrow$ Vietnamese translation task compared to the baseline systems.

(3) Multilingual fine-tuning systems. The multilingual system is fine-tuned from the baseline for further 15 epochs with an initial learning rate of 0.05. We see the improvements of +1.43 and

[1] https://www.ted.com/
[2] https://www.statmt.org/europarl
[3] https://github.com/quanpn90/NMTGMinor
[4] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

Datasets	Systems	dev	test
English → Vietnamese	Bilingual Baseline	31.74	35.13
	Multilingual	31.66 (-0.08)	**36.18 (+1.05)**
	Multilingual + fine-tuning	**31.88 (+0.14)**	**36.56 (+1.43)**
	Multilingual + fine-tuning with similarity	**31.93 (+0.19)**	**36.75 (+1.62)**
	Multilingual + fine-tuning with updated embedding	**32.11 (+0.37)**	**36.74 (+1.61)**
	Multilingual + mixing pseudo bilingual data	30.86 (-0.88)	35.09 (-0.04)
French → Vietnamese	Bilingual Baseline	23.07	23.03
	Multilingual	**24.49 (+1.42)**	**24.22 (+1.19)**
	Multilingual + fine-tuning	**24.51 (+1.44)**	**24.86 (+1.83)**
	Multilingual + fine-tuning with similarity	**24.37 (+1.30)**	**24.70 (+1.63)**
	Multilingual + fine-tuning with updated embedding	**24.60 (+1.53)**	**24.96 (+1.93)**
	Multilingual + mixing pseudo bilingual data	**25.59 (+2.52)**	**25.57 (+2.54)**
	Pseudo bilingual data translation	19.00	18.71

Table 2: The results of our MT systems are measured in BLEU. We evaluate the best model for the baseline systems and the average scores on the five best models for the multilingual and pseudo systems.

+1.83 BLEU points on both translation tasks, respectively.

(4) Multilingual fine-tuning with similarity systems. The systems from (2) are fine-tuned with the strategy mentioned in section 4.1 using the modified framework. We obtained a bigger gain of +1.62 BLEU points on the English → Vietnamese translation task whilst the French → Vietnamese translation task has achieved a lower improvement than the systems in (3). We show that the English → Vietnamese translation task has more advantages when rare tokens from French are replaced by similar tokens in the multilingual space. In the future, we would attempt the inverse replacement.

(5) Multilingual fine-tuning with updated embedding systems. We use the modified framework to fine-tune the systems in (2) with the method mentioned in section 4.2. The greater improvements can be found at +1.61 and +1.93 on both translation tasks compared to the systems which do not use our methods.

(6) Multilingual with mixing of pseudo bilingual data. We use 400K synthetic bilingual sentence pairs for each of the language pairs: English-Vietnamese and French-Vietnamese. We train the multilingual NMT system on a mix of pseudo and real bilingual data mentioned in section 5 for 50 epochs. And then it is fine-tuned on the actual parallel data for 20 epochs. We observed a bigger improvement of **+2.54** BLEU points on the French → Vietnamese system while the English → Vietnamese system has achieved less improvement compared to previous systems. We speculate that the English → Vietnamese translation task may be affected by the French → Vietnamese pseudo bilingual data. In future work, we would leverage the data selection methods in order to equip better synthetic data for our systems.

(7) Pseudo bilingual data translation. We train the French → Vietnamese NMT system relied on only 1.2 thousands pseudo bilingual data mentioned in section 5 for 26 epochs. We achieve 18.71 BLEU points on the averaged model from our five best models. Thus, we can generate synthetic parallel data for a low-resource language pair from another language pair with a bigger bilingual resource.

7 Related Work

Due to the unavailability of the parallel data for low-resource language pairs or zero-shot translation, previous works focus on the task to have more data such as leveraging multilingual translation (Ha et al., 2016, 2017; Wang et al., 2018; Gu et al., 2019; Aharoni et al., 2019) or using monolingual data with back-translation, self-learning (Sennrich et al., 2015; Zhang and Zong, 2016; Wu et al., 2019) or mix-source (Ha et al., 2016) technique.

For leveraging multilingual translation, (Ha et al., 2016) added language code and target forcing in order to learn the shared representations of the source words and specify the target words. (Wang et al., 2018) demonstrated a one-to-many multilingual MT with three different strategies which modify their architecture. (Gu et al., 2019) built many-to-one multilingual MT systems by adding a layer to transform the source embeddings and representation into a universal space to augment the translation of low resource language, which is similar to ours. (Aharoni et al., 2019) implemented a massive many-to-many multilingual system, employing many low-resource language pairs. All of the mentioned works have shown substantial improvements in low-resource translation, however,

they are less correlative to our translation tasks.

Although multilingual MT equips a shared space with many advantages, rare word translation is still the issue that needs to be considered. The task of dealing with rare words has been mentioned in previous works. (Luong et al., 2015) copied words from source sentences by words from target sentences after the translation using a bilingual dictionary. (Li et al., 2016) and (Trieu et al., 2016) learned word similarity from monolingual data to improve their systems. Our approach is similar to these works, but we only learn similarity from the shared multilingual space of MT systems. (Ngo et al., 2019) addressed the rare word problem by using the synonyms from WordNet.

(Nguyen and Chiang, 2017) and (Ngo et al., 2019) presented different solutions to solve rare word situation by transforming the embeddings during the training of their RNN-based architecture. Those solutions cannot be applied to the transformer architecture. In (Gu et al., 2019), the embeddings of rare tokens and universal tokens are jointly learned through a plus parameter while we only add a scalar value to the embeddings.

Monolingual data is used to generate synthetic bilingual data in sparsity data issues. (Sennrich et al., 2015) proposed back-translation method that uses a backward model to get the source data from the monolingual target data. In contrast, (Zhang and Zong, 2016) shown the self-learning technique by employing a forward model to translate monolingual source data into the target data. (Wu et al., 2019) incorporated both mentioned techniques into their NMT systems. Monolingual data is also demonstrated its efficiency in unsupervised machine translation(Lample et al., 2018) or in zero-shot multilingual NMT (Siddhant et al., 2020; Ha et al., 2017). In our work, we use the self-learning method to produce pseudo bilingual data, and it is then used to train our low-resource multilingual NMT systems.

8 Conclusion and Future Work

We have built multilingual MT systems for two low-resource language pairs: English-Vietnamese and French-Vietnamese, and proposed two approaches to tackle rare word translation. We show that our approaches bring significant improvements to our MT systems. We find that the pseudo bilingual can furthermore enhance a multilingual NMT system in case of French → Vietnamese translation task. In the future, we would like to use more language pairs in our systems and to combine proposed methods in order to evaluate the effectiveness of our MT systems.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *CoRR*, abs/1903.00089.

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

M Cettolo, J Niehues, S Stüker, L Bentivogli, R Cattoni, and M Federico. 2016. The IWSLT 2016 Evaluation Campaign. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, WA, USA.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *CoRR*, abs/1808.09381.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. *CoRR*, abs/1906.01181.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2017. Effective Strategies in Zero-Shot Neural Machine Translation.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *IJCAI*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

Processing (Volume 1: Long Papers), pages 11–19, Beijing, China. Association for Computational Linguistics.

Thi-Vinh Ngo, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. 2019. Overcoming the rare word problem for low-resource language pairs in neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 207–214, Hong Kong, China. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Improving lexical choice in neural machine translation. *Proceedings of NAACL-HLT 2018*, pages 334–343.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Association for Computational Linguistics (ACL 2016)*.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. *CoRR*, abs/1905.11901.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Hai-Long Trieu, Le-Minh Nguyen, and Phuong-Thai Nguyen. 2016. Dealing with out-of-vocabulary problem in sentence alignment using word similarity. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 259–266, Seoul, South Korea.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. pages 2955–2960.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for 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 4207–4216, Hong Kong, China. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.