Back-translation for Large-Scale Multilingual Machine Translation

Baohao Liao Shahram Khadivi Sanjika Hewavitharana
eBay Inc.
{baliao|skhadivi|shewavitharana}@ebay.com

Abstract

This paper illustrates our approach to the shared task on large-scale multilingual machine translation in the sixth conference on machine translation (WMT-21). In this work, we aim to build a single multilingual translation system with a hypothesis that a universal cross-language representation leads to better multilingual translation performance. We extend the exploration of different back-translation methods from bilingual translation to multilingual translation. Better performance is obtained by the constrained sampling method, which is different from the finding of the bilingual translation. Besides, we also explore the effect of vocabularies and the amount of synthetic data. Surprisingly, the smaller size of vocabularies perform better, and the extensive monolingual English data offers a modest improvement. We submitted to both the small tasks and achieve the second place. The code and trained models are available at https://github.com/BaohaoLiao/multiback.

1 Introduction

Bilingual neural machine translation (NMT) systems have achieved decent performance with the help of Transformer (Vaswani et al., 2017). One of the most exciting recent trends in NMT is training a single system on multiple languages at once (Johnson et al., 2017b; Aharoni et al., 2019a; Zhang et al., 2020; Fan et al., 2020). This is a powerful paradigm for two reasons: simplifying system development and deployment, and improving the translation quality on low-resource language pairs by transferring similar knowledge from high-resource languages.

This paper describes our experiments on the task of large-scale multilingual machine translation in WMT-21. We primarily focus on the small tasks, especially on Small Task 2 which has a small amount of training data. Small Task 1 contains five Central/East European languages and English, having 30 translation directions. Similarly, Small Task 2 contains five South East Asian languages and English, also having 30 translation directions.

In this work, we mainly concentrate on different back-translation methods (Sennrich et al., 2016a; Edunov et al., 2018; Graça et al., 2019) for multilingual machine translation, including beam search and other sampling methods. Along with it, we also explore the effect of different sizes of vocabularies and the effect of various amounts of synthetic data. On this large-scale multilingual machine translation task, we achieved the second place for both small tasks, obtaining 34.96 and 33.34 average sp-BLEU scores (Goyal et al., 2021) on the hidden test set for the Small Task 1 and 2, respectively.

2 Related Work

Multilingual Neural Machine Translation has received increasing attention recently. Since Dong et al. (2015) extended the traditional bilingual NMT to one-to-many translation, there has been a massive increase in work on MT systems that involve more than two languages (Dabre et al., 2017; Choi et al., 2018; Chu and Dabre, 2019). The recent research on multilingual NMT can be split into two directions: developing language specific components (Kim et al., 2019; Escolano et al., 2020) and training a single model with extensive training data, including parallel and monolingual data (Fan et al., 2020). Here, we continue to explore the second research direction, trying to built a single multilingual NMT model for simple industrial deployment.

Back-translation (Sennrich et al., 2016a) has been proven as a powerful technique to leverage monolingual data for improving low-resource language pairs. Edunov et al. (2018) and Graça et al. (2019) explore different sampling methods for bilingual back-translation, including beam search, constrained and unconstrained sampling. Constrained sampling randomly predicts the next word

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 418–424
November 10–11, 2021. ©2021 Association for Computational Linguistics

within some candidates that have higher prediction probability. And unconstrained sampling randomly predicts the next words from the whole vocabulary without caring for the output distribution. In this paper, we extend their exploration to the realm of multilingualism, where similar languages affect the results.

3 Experimental Setup

3.1 Data

The organizer offers parallel and monolingual data for Small Task 1 and 2. Table 1 shows the size of the data in terms of the number of sentences for each language. There are five extra sets for evaluation, i.e. dev, devtest, hidden dev, hidden devtest and test sets. The dev set with 997 parallel sentences among all language pairs and the devtest set with 1,012 parallel sentences are public. Whereas, the hidden dev and hidden devtest sets are invisible to the participants and used for the first submission period. The hidden test set is also invisible and used for the final ranking.

Pre-processing is done by a regular Moses toolkit (Koehn et al., 2007) pipeline that involves tokenization, byte pair encoding and removing long sentences. We borrow the 256K vocabularies from the organizer's pretrained model and the 128K vocabularies from M2M_100 (Fan et al., 2021), one shared vocabularies among all languages. Our submissions only use the 256K vocabularies, and the 128K vocabularies is used for ablation experiments.

We also perform back-translation on the monolingual data, and only accept the synthetic sentence pair whose length is less than 250 words, and whose length ratio between the source and target sentence length is less than 1.8. In order to balance the volume across different languages, we apply temperature sampling $\tilde{D}_i = (D_i / \sum_j D_j)^{1/T}$ with $T = 5$ over the dataset, where D_i is the number of sentences in the i_{th} language.

3.2 Model

All our models are built using the fairseq implementation (Ott et al., 2019) of the Transformer architecture (Vaswani et al., 2017). Multilingual models are built using the same technique as Johnson et al. (2017a) and Aharoni et al. (2019b), namely adding a language label to the target sentence.

We apply three types of architectures, i.e. *Trans_small*, *Trans_base* and *Trans_big*. The detailed settings of these architectures are shown in Table 2. The parameters of all architectures are in the half-precision floating-point format.

Small Task 1		**Small Task 2**	
Language	**#sent.**	**Language**	**#sent.**
en-et	35.7M	en-id	54.1M
en-hr	63.7M	en-jv	3.0M
en-hu	83.9M	en-ms	13.4M
en-mk	2.7M	en-ta	2.1M
en-sr	48.3M	en-tl	13.6M
et-hr	13.6M	id-jv	780.1K
et-hu	21.5M	id-ms	4.9M
et-mk	3.1M	id-ta	500.8K
et-sr	11.3M	id-tl	2.7M
hr-hu	31.2M	jv-ms	434.7K
hr-mk	4.4M	jv-ta	66.0K
hr-sr	28.4M	jv-tl	817.1K
hu-mk	4.1M	ms-ta	372.6K
hu-sr	31.2M	ms-tl	1.4M
mk-sr	4.2M	ta-tl	563.3K
en	126.4M	en	126.4M
et	3.0M	id	5.5M
hr	3.1M	jv	405.8K
hu	9.2M	ms	1.9M
mk	1.9M	ta	2.1M
sr	4.7M	tl	414.1K

Table 1: Number of sentences of the parallel and monolingual data used for two small tasks. The monolingual English data for the two small tasks are the same.

All our submissions on the shared task leaderboard are *Trans_base*, due to the memory and time limit of the evaluation system. *Trans_small* is mainly used for the ablation experiments. And the pretrained *Trans_big* from M2M_100 (Fan et al., 2021) is finetuned on the parallel corpus to generate high-quality synthetic sentences.

3.3 Optimization and Evaluation

The following hyper-parameter configuration is used: Adam optimizer with $\beta_1 = 0.90$, $\beta_2 = 0.98$, a weight-decay of 0.0001, the label smoothed cross-entropy criterion with a label smoothing of 0.1, an initial learning rate of 0.0003 with the inverse square root lr-scheduler and warmup updates of 2,500 steps. The batch size (the number of tokens) is 4096×32 for *Trans_small*, and 2048×64 for *Trans_base* and *Trans_big*.

For ablation experiments, we continue to train the pretrained *Trans_small* offered by the organizer

Model	Trans_small	Trans_base	Trans_big
#vocabularies	256K	256K	128K
Word representation size	512	1,024	1,024
Feed-forward layer dimension	2,048	4,096	8,192
#prenormed encoder/ decoder layer	6	12	24
#attention head	16	16	16
Dropout rate	0.1	0.1	0.1
Layer dropout rate	0.05	0.05	0.05
#parameters	175M	615M	1.2B

Table 2: Settings of different pretrained models. Pretrained *Trans_small* and *Trans_base* are provided by the organizer. And pretrained *Trans_big* is from Fan et al. (2021).

on the given parallel dataset for one epoch. When combining both parallel and synthetic data, we further train the model finetuned on the parallel data for another one epoch. For the final submissions, we train a pretrained *Trans_base* for two epochs instead of one epoch. Pretrained *Trans_big* from M2M_100 is only further trained on parallel data for two epochs to generate high-quality synthetic data. Even though we only train these models for a few epochs, they seems converged quite well according to the spBLEU curve during validation.

The model is validated every 3,000 steps on the dev set and saved. We use the beam search with a beam size of five, and stop translation when $l_{tgt} = 1.5 * l_{src} + 20$, where l_{src} and l_{tgt} are the source and target sentence length, respectively. The evaluation metric is BLEU based on sentence piece tokenization (spBLEU) (Goyal et al., 2021). We submit the average checkpoint of the last 15 checkpoints to the evaluation system. While for the ablation experiment, we use the best performed model on the dev set.

4 Results

4.1 The Role of Vocabularies

There are two pretrained vocabularies, the one with the size of 256K from the organizer and the one with the size of 128K from M2M_100 (Fan et al., 2021). To evaluate which vocabulary is the better one, we train two *Trans_smalls* with these two vocabularies from scratch on the parallel data of Small Task 2 for five epochs. To make the parameter sizes of these two models comparable, we set the following hyper-parameter for the model with the 128K vocabularies: 5 pre-normed encoder and decoder layers with a word representation size of 768 and a feed-forward layer dimension of 3072, resulting to 181M parameters. The other settings stay the same with *Trans_small* (with 256K vocabularies).

Model	Ave. spBLEU
128K *Trans_small* (scratch)	23.14
256K *Trans_small* (scratch)	21.65
256K *Trans_small* (pretrained)	23.72

Table 3: Average spBLEU on the devtest set of Small Task 2 for the models with different vocabularies.

Model	Ave. spBLEU
1st finetuned on parallel data	28.27
2nd finetuned on synthetic data	32.16
3rd finetuned on synthetic data	33.01

Table 4: Average spBLEU on the devtest set of Small Task 2 for *Trans_base* on different finetuning steps. These three models are iteratively trained. *Trans_base* is first finetuned on the parallel data, and then finetuned on the combination of the parallel data and the synthetic data generated by *Trans_big*, and finally finetuned on the combination of the parallel data and the synthetic data generated by the 2nd step *Trans_base*.

Table 3 shows the performance with different vocabularies. It is obvious that the 128K vocabulary outperforms the 256K vocabulary, 23.14 vs 21.65 spBLEU. However, if we finetune the pretrained *Trans_small* with the 256K vocabulary, 0.58 score improvement is achieved compared to the 128K *Trans_small*. In a word, 128K vocabulary is a better choice for training from scratch, while pretrained model offers us more gain.

Model	Ave. spBLEU
1st finetuned on parallel data	32.46
2nd finetuned on synthetic data	34.73

Table 5: Average spBLEU on the devtest set of Small Task 1 for *Trans_base* on different steps. These two models are iteratively trained. *Trans_base* is first finetuned on the parallel data, and then finetuned on the combination of the parallel data and the synthetic data generated by the previous step *Trans_base*.

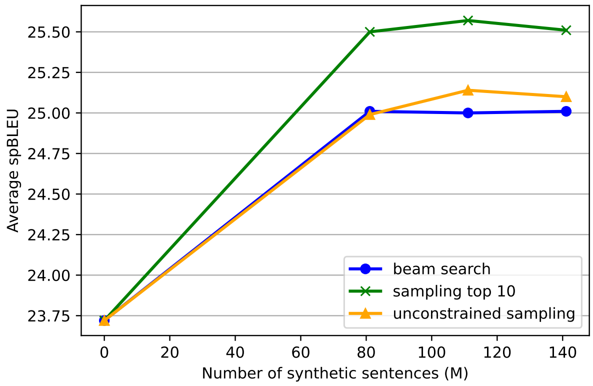

Figure 1: Average spBLEU on the devtest set of Small Task 2 for different back-translation methods with various amount of synthetic data. 80M synthetic data covers only 6M monolingual English data and all other monolingual data. We increase the amount of monolingual English data with a interval of 6M for the last two experiments.

4.2 Different Back-translation Methods

Similar to Edunov et al. (2018), we explore three types of back-translation methods, i.e. beam search with the beam size of five (Sennrich et al., 2016a), unconstrained sampling (Edunov et al., 2018) and sampling constrained to the most 10 likely words (Graves, 2013; Ott et al., 2018; Fan et al., 2018). Unconstrained sampling predicts the next word from the whole vocabulary without caring for the model distribution. Whereas constrained sampling predicts the next words within some candidates that have the highest prediction probabilities. Both constrained and unconstrained sampling can be considered as adding uncertainty to the greedy search.

Figure 1 shows the back-translation results on the devtest set of Small Task 2. We combine three different amount of synthetic data and parallel data to further train our *Trans_small*s after finetuned on parallel data. 80M synthetic sentences cover only 6M monolingual English data and all other monolingual data. In addition to the 80M synthetic sentences, we further increase the amount of monolingual English data to verify the model performance with respect to the amount of synthetic English data on the target side. The reason for this implementation is there are too many monolingual English sentences compared to other languages. We try to check whether it is necessary to use all monolingual English sentences.

As seen in Figure 1, little improvement is obtained with increasing the number of monolingual English sentences after 6M. Besides, in contrast to the results in Edunov et al. (2018) where the unconstrained sampling offers the best performance among these three methods, the constrained sampling method gives us the best score.

Beam search is the worst among these three methods. We hypothesize this is because beam search focuses only on the high probability words, while both constrained sampling and unconstrained sampling methods offer rich translations on the source side. With the diverse synthetic data generated from the sampling methods, model can be trained with more generalization.

In contrast to the bilingual translation (English-German) in Edunov et al. (2018) where unconstrained sampling outperforms constrained sampling, multilingual translation of Small Task 2 contains similar languages. We argue that unconstrained sampling might result in generating synthetic sentences with a mix of similar languages, which damages the quality of synthetic data, while constrained sampling gives us some restriction, to some extent avoiding the mix of different languages.

The reason for the slight effect of the synthetic English (on the source side) data after 6M might be that English is dissimilar to the other five South East Asian languages. Less similar knowledge could be transferred from this synthetic English (on the source side) data to other languages.

4.3 Final Submissions

Section 4.1 suggests us to employ a pretrained model with the 128K vocabulary. M2M_100 (Fan et al., 2021) offers multiple pretrained models with the 128K vocabularies [1]. Their sizes are 418M, 1.2B and 12B, respectively. Considering our limited GPU budget, we finetune the 1.2B model, i.e.

[1] https://github.com/pytorch/fairseq/tree/master/examples/m2m_100

Small Task	devtest	hidden dev	hidden devtest	hidden test
#1	34.73	35.12	35.39	34.96
#2	33.01	33.74	33.51	33.34

Table 6: Average spBLEU on different test sets for both small tasks. The hidden sets are invisible to the participants. The final ranking is based on the model performance on the hidden test set.

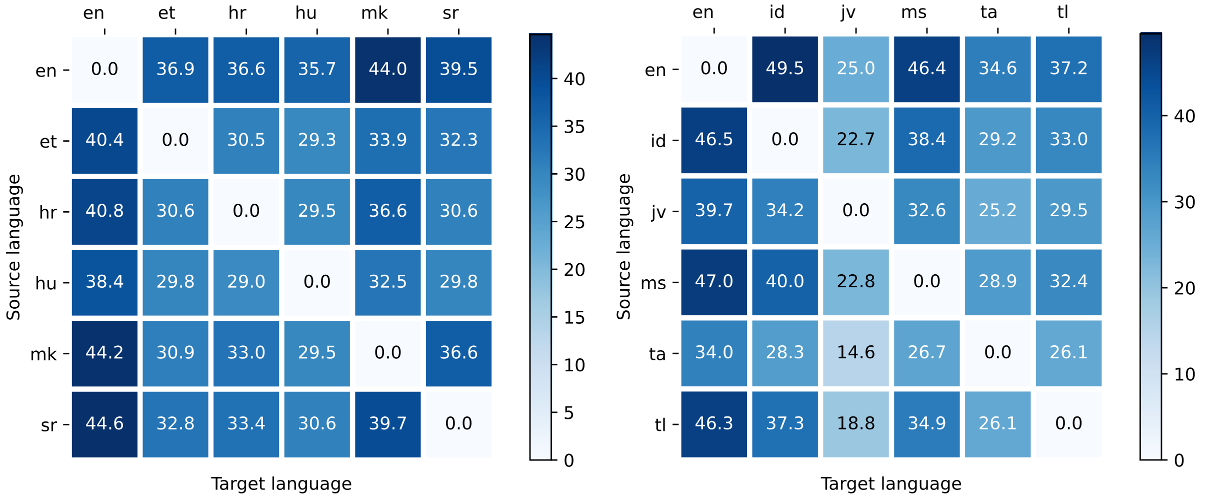

Figure 2: The spBLEU scores of different language pairs for both small tasks on the devtest set from our final submissions.

Trans_big, on parallel data of Small Task 2, obtaining 28.78 spBLEU on the devtest set. Whereas, training a *Trans_base* on the same data only provides 28.23 spBLEU. Even though *Trans_big* outperforms *Trans_base*, we only train it for generating high-quality synthetic data, since it is too large for the evaluation system.

Section 4.2 advises us to use the constrained sampling method on partial monolingual English data. With the constrained sampling method, we generate synthetic sentences with *Trans_big* that is first finetuned on the parallel data. Instead of using all monolingual English data, we synthesize en-id, en-jv, en-ms, en-ta and en-tl with all, 15M, 60M, 10M and 60M monolingual English sentences, respectively, a ratio of about 5 : 1 between the number of parallel sentences and synthetic sentences if there are enough monolingual data.

Table 4 shows the results for iterative finetuning. Except for finetuing *Trans_base* on the combination of the parallel data and the synthetic data generated by *Trans_big*, we use the finetuned *Trans_base* to generate the synthetic data secondly and finetune it again. Finally, it offers us 33.01 spBLEU on the devtest set for Small Task 2.

Due to time and resource limit, we only conduct one trial on Small Task 1. We first finetune the pretrained *Trans_base* on parallel data. Then we use this *Trans_base* to generate synthetic data with only 20M monolingual English sentences and all other monolingual sentences. Table 5 shows the corresponding results. Different with Small Task 2, large amount of monolingual English data might be helpful for Small Task 1, since Central/East European languages are more similar to English than Asian languages. Finally, We leave this exploration to the future work.

Table 6 summarizes the results of our submissions on different evaluation sets for both small tasks. And Figure 2 lists the spBLEU scores for all language pairs of both small tasks on the devtest set. Finally, our submissions achieve the second place for both small tasks.

5 Conclusion

We demonstrate that a pretrained model with the smaller size of vocabularies is a better choice. Because of the memory and time limit of the evaluation system, we can only apply a 1.2B model with the smaller vocabularies to generate high-quality synthetic data. Besides, we have a different observation than previous research for bilingual back-translation: the constrained sampling method performs the best among all three back-translation methods, including the beam search and the unconstrained sampling. Finally, we also show that extensive monolingual English data offers a modest improvement. Combining these three findings, we iteratively train our models on partial high-quality synthetic data, achieving the second place for both small tasks.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019a. 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.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019b. 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 (NAACL-HLT 2019)*, page 3874–3884.

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).

Chenhui Chu and Raj Dabre. 2019. Multilingual multi-domain adaptation approaches for neural machine translation. *CoRR*, abs/1906.07978.

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.

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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2020. Training multilingual machine translation by alternately freezing language-specific encoders-decoders. *CoRR*, abs/2006.01594.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

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

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017b. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965. PMLR.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Conference of the Association for Computational Linguistics (ACL)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.