Transfer Learning for Related Languages: Submissions to the WMT20 Similar Language Translation Task

Lovish Madaan, Soumya Sharma, Parag Singla
Indian Institute of Technology, Delhi
{lovish97, soumyasharma98}@gmail.com, parags@cse.iitd.ac.in

Abstract

In this paper, we describe IIT Delhi's submissions to the WMT 2020 task on Similar Language Translation for four language directions: Hindi ↔ Marathi and Spanish ↔ Portuguese. We try out three different model settings for the translation task and select our primary and contrastive submissions on the basis of performance of these three models. For our best submissions, we fine-tune the mBART model (Liu et al., 2020) on the parallel data provided for the task. The pre-training is done using self-supervised objectives on a large amount of monolingual data for many languages. Overall, our models are ranked in the top four of all systems for the submitted language pairs, with first rank in Spanish → Portuguese.

1 Introduction

Machine Translation (MT) is currently tackled using rule-based methods (RBMT) (Charoenpornsawat et al., 2002), phrase-based statistical methods (SMT) (Koehn et al., 2003) and neural methods (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017).

NMT has achieved high translation quality for several language pairs (Bojar et al., 2018; Barrault et al., 2019), but this level of performance usually requires large amounts of aligned data in the order of millions of sentence pairs. For low and medium resource languages, SMT performs better than NMT (Koehn and Knowles, 2017; Sennrich and Zhang, 2019). SMT also shows better performance when there is a domain mismatch between the train and test datasets, which is typical of low and medium resource language pairs.

In these settings, NMT performance can be boosted by leveraging additional monolingual data to enforce various types of constraints or increasing the training data using back-translation. These methods can be particularly helpful if the source and target languages in MT are closely related and share language structure and alphabet. Recently, pre-training methods for sequence-to-sequence (seq2seq) models have been introduced like MASS (Song et al., 2019a), XLM (Conneau and Lample, 2019), BART (Lewis et al., 2019), and mBART (Liu et al., 2020). These methods show significant gains in downstream tasks like NMT, summarization, natural language inference (NLI), etc. In this paper, we focus on the transfer learning capabilities in NMT for the task of translation between related languages where parallel data is scarce.

IIT Delhi participated in the WMT 2020 Shared task on Similar Language Translation for four language directions: `Hindi (hi)` ↔ `Marathi (mr)` and `Spanish (es)` ↔ `Portuguese (pt)`. The first language pair is low resource and second is medium resource in terms of the parallel data available for the task. Refer to Table 2 for the classification.

We fine-tuned the pre-trained mBART model (Liu et al., 2020) on the parallel data provided for the task. mBART gives better performance than SMT models even when the parallel data is very limited. mBART is pre-trained on 25 languages, which contain Hindi and Spanish, but not Marathi and Portuguese. mBART is able to leverage transfer learning capabilities even for those languages that are originally not present during the pre-training phase. The fine-tuned mBART architecture forms our best submissions for both language pairs: `hi` ↔ `mr` and `es` ↔ `pt`. The rankings obtained by us in each of the language directions are listed in Table 1. Our findings are in line with earlier observations in the literature where transfer learning techniques have been shown to significantly boost NMT performance.

The rest of the paper is organized as follows: Section 2 provides the background and related work for low/medium resource NMT. Section 3 gives an

Proceedings of the 5th Conference on Machine Translation (WMT), pages 402–408
Online, November 19–20, 2020.

Direction	BLEU	Rank
hi → mr	15.14	4
mr → hi	24.53	2
es → pt	32.69	1
pt → es	32.84	2

Table 1: BLEU scores on the test set provided for the task and system rankings according to the automatic evaluation metrics.

overview of the systems tried. In Section 4, we present the experiments and training pipeline setup. The results and analysis are detailed in Section 5. We finally conclude in Section 6.

2 Background

SMT is tackled by building a phrase table from the aligned parallel data. The target side translation is then generated by matching the most appropriate phrases in the source sentence conditioned on the target side language model along with a reordering model (Koehn et al., 2003).

NMT is modeled using Encoder-Decoder models (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015), with the Transformer model (Vaswani et al., 2017) achieving state-of-the-art on many MT problems. But these models' reliance on large aligned parallel data for the source and target languages makes them unsuitable for low/medium resource language pairs (Koehn and Knowles, 2017). Some of the previous works in these settings to improve NMT performance are described below:

2.1 Multilingual NMT

Instead of using only two languages (source and target) for training an NMT model, using multiple languages has been shown to help in low resource scenarios. For example, it might be the case that a certain pair of languages have very little parallel data between them, but there exists a third language with abundant parallel data with the original two languages. This third language acts as a pivot and helps in improving NMT between the two languages (Aharoni et al., 2019; Gu et al., 2018; Liu et al., 2020; Zhang et al., 2020).

2.2 Back-Translation

Back-Translation (Sennrich et al., 2016; Edunov et al., 2018; Hoang et al., 2018) increases the amount of training data by using monolingual corpus along with partially-trained NMT models on the limited parallel data. Pseudo-parallel corpus for each direction is first obtained by generating the translations of the monolingual data for each language using the partially-trained MT models on the limited parallel data. Using these pseudo-parallel corpora, the partially-trained NMT models are then trained further for some number of steps. In this way, millions of pseudo-parallel sentence pairs can be generated to improve NMT models because of the abundance of monolingual data. Another version of using back-translation is the copying mechanism. Currey et al. (2017) proposes to copy the target side monolingual data on the source side to create additional data without modifying the training regimen for NMT. This helps the model to generate fluent translations.

2.3 Pre-trained Language Models

For NMT, the first step is the random initialization of model weights in both the encoder and decoder. Instead of random initialization, NMT models can be initialized by pre-training parts of the model (Conneau and Lample, 2019; Edunov et al., 2019), or pre-training the complete seq2seq model (Ramachandran et al., 2017; Song et al., 2019b; Liu et al., 2020). These pre-training methods leverage different kinds of masking techniques and the pre-training objective is to predict these masked tokens, similar to BERT (Devlin et al., 2019). Denoising auto-encoding can also be used where a sentence is corrupted by various noising techniques and the pre-training objective is to generate the original uncorrupted sentence as in BART (Lewis et al., 2019) and mBART (Liu et al., 2020).

2.4 Incorporating Linguistic Information in NMT

There also have been works to improve low/medium resource NMT by adding linguistic information either using data augmentation (Currey and Heafield, 2019), subword embedding augmentation (Sennrich and Haddow, 2016) , or architectural changes (Eriguchi et al., 2017). This helps the model to not only learn the alignment between source and target language spaces, but also syntax structure like dependency parse, part of speech, etc. This helps in making the target side translations more fluent and conforming to the structure of the language. We do not explore this direction in this paper.

3 System Overview

We experimented with three different settings for hi ↔ mr as listed below.

SMT This phrase-based system leverages both monolingual and parallel data provided for the task. We use Moses (Koehn et al., 2007) for training the SMT systems.

NMT (Transformer) For this, we used the standard Transformer large architecture from Vaswani et al. (2017) for training on the parallel data provided for the task.

NMT (mBART) mBART (Liu et al., 2020) is a large Transformer pre-trained on monolingual data for 25 languages. The pre-training objective for mBART is seq2seq de-noising for natural text as in BART (Lewis et al., 2019). mBART provides a general-purpose pre-trained Transformer for any downstream task. It has been shown to give significant improvements over the random initialization for NMT and is the current state-of-the-art for many low resource language pairs.

Implementation Details mBART uses a shared subword vocabulary of 250K tokens for all the 25 languages present in the pre-training. We use the same vocabulary for Marathi and Portuguese also, even though they were not used during the pre-training phase. Marathi shares its subword vocabulary with languages like Hindi and Nepali in mBART, and Portuguese shares with Spanish, Italian and other European languages present in mBART. The percentage of unknown tokens [UNK] in Marathi and Portuguese parallel datasets is less than 0.003% when using the shared mBART vocabulary.

Additionally, the mBART architecture requires language specific token at the end of each input sequence to provide the language specific context for the decoder. Since Marathi and Portuguese were not present during the pre-training phase, we use the token corresponding to the second most related language present in mBART pre-training for specifying the context at the time of decoding in each case. For Marathi, we used the Nepali language token and for Portuguese, we used the Italian language token. We could not use Spanish language token for Portuguese because we are doing translations to and from Spanish.

	train	valid	test
hi ↔ mr	43,274	1,411	1,941
es ↔ pt	3,472,860	1,283	1,495

Table 2: Dataset statistics. First is low resource pair (# train < 1 Million) and second is medium resource (1 Million < # train < 10 Million).

Model	**hi - mr**	
	←	→
SMT	18.74	14.91
mBART	**24.53**	**15.14**

Table 3: BLEU scores on Hindi ↔ Marathi on the test set for our primary and contrastive submissions.

4 Experiments

We use hi ↔ mr and es ↔ pt language pairs for our experiments.

4.1 Datasets & Preprocessing

Because of the constrained nature of the shared task, we only use the parallel data provided for this task. We removed the empty instances for both language pairs (< 2000 instances). For es ↔ pt, we do not use 'WikiTitles v2' part of the parallel data for training because of very short sentences in the dataset. The cleaned parallel dataset statistics are provided in Table 2.

Preprocessing We use sentence piece tokenization (Kudo and Richardson, 2018) for generating the source and target sequences for the NMT architectures. For the standard Transformer, we train a sentence piece model using 40K subword tokens for hi ↔ mr. For mBART, we use Liu et al. (2020)'s pre-trained[1] sentence piece model comprising of 250K subword tokens as the vocabulary.

For the SMT model on hi ↔ mr, we also use the monolingual data provided for this task. We extract 5 Million monolingual sentences each for Hindi and Marathi after deduplication and use this set for training the language models. We use Moses (Koehn et al., 2007) for all tokenization / detokenization scripts.

[1] https://github.com/pytorch/fairseq/blob/master/examples/mbart/README.md

Submission	hi - mr		es - pt	
	←	→	←	→
IIT Delhi (ours)	**24.53**	15.14	32.84	**32.69**
Rank 1	24.53	18.26	33.82	32.69

Table 4: Hindi - Marathi and Spanish - Portuguese BLEU scores on the test dataset of the Similar Language Translation Task. Our submission scores are bolded when they match the first ranked submission.

4.2 Model Architectures & Training

SMT We generate a phrase table for the SMT model using the code provided by Lample et al. (2018). We used Moses (Koehn et al., 2007) and Giza++ with standard settings to train the SMT model in both directions.

NMT (Transformer) We use the large Transformer from Vaswani et al. (2017) with 8 encoder and decoder layers and replicate all the parameters from Ott et al. (2018). The number of parameters in the model are approximately 248 Million and it takes ∼26 hours on 4 Nvidia V100 (32 GB) GPUs.

NMT (mBART) For this, we use 12 Transformer encoder and decoder layers, with total number of model parameters ∼611 Million. We use the pre-trained mBART for initializing the model weights. We follow the recommendations of Liu et al. (2020) for the hyperparameter settings. We stop the training after 25K gradient updates for the model. These updates take ∼35 hours on 4 Nvidia V100 (32 GB) GPUs.

4.3 Evaluation

We use case-insensitive BLEU scores (Papineni et al., 2002) calculated using sacreBLEU[2] (Post, 2018). These scores are calculated on the validation set to decide our primary and contrastive submissions. For evaluating performance on the test set, the organizers use BLEU, TER (Snover et al., 2006), and RIBES (Isozaki et al., 2010).

5 Results and Analysis

Results Table 3 shows our results on the test set for our primary and contrastive submissions. We observed the performance of our three model settings on the validation set, and we selected the mBART model as our primary submission and SMT model as the contrastive submission for `hi` ↔ `mr`. Similarly, the mBART model forms our primary submission for `es` ↔ `pt`. Table 4 lists our final results on this shared task. We also list the BLEU scores for the submission that got first rank in each of the language directions. Since the test sets were hidden at the time of submission, we do not report our numbers on the standard Transformer architecture.

Analysis Even though Marathi and Portuguese are not present during the pre-training phase of mBART, fine-tuning on these languages provides significant boosts over SMT and standard Transformer. This shows that some level of language independent multilingual embeddings are present in the pre-trained model weights which can be exploited for the transfer task.

6 Discussion and Conclusion

We have participated in the Similar Language Translation task on four language directions. We have shown that pre-trained models can help in low and medium resource NMT. Our best system uses the pre-trained mBART model (Liu et al., 2020) and fine-tunes on the parallel data provided for the specific translation task. Our results demonstrate that pre-training can help even when the language used for fine-tuning is not present during pre-training.

One direction of future work is to add linguistic information during the pre-training phase to get more fluent translations. When this information is not available directly (especially for low resource languages), pre-training on a related high resource language with syntax information can help low resource languages also.

Acknowledgments

We thank the IIT Delhi HPC facility[3] for the computational resources. We are also thankful to Ganesh Ramakrishnan and Pawan Goyal for initial discussions on the project. Parag Singla is supported

[2]Signature: BLEU + case.mixed + numrefs.1 + smooth.exp + tok.13a + version.1.3.1

[3]http://supercomputing.iitd.ac.in/

by the DARPA Explainable Artificial Intelligence (XAI) Program with number N66001-17-2-4032, Visvesvaraya Young Faculty Fellowships by Govt. of India and IBM SUR awards. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of the funding agencies.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3874–3884. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Loïc Barrault, Ondrej 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, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 1–61. Association for Computational Linguistics.

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 272–303. Association for Computational Linguistics.

Paisarn Charoenpornsawat, Virach Sornlertlamvanich, and Thatsanee Charoenporn. 2002. Improving translation quality of rule-based machine translation. In *COLING-02: Machine Translation in Asia*.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 148–156. Association for Computational Linguistics.

Anna Currey and Kenneth Heafield. 2019. Incorporating source syntax into transformer-based neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 1: Research Papers*, pages 24–33. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4052–4059. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. Association for Computational Linguistics.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 72–78. Association for Computational Linguistics.

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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 344–354. Association for Computational Linguistics.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 18–24. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 944–952. ACL.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

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, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Prajit Ramachandran, Peter J. Liu, and Quoc V. Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 383–391. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 211–221. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation.

In *Proceedings of Association for Machine Translation in the Americas, 2006*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019a. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019b. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *CoRR*, abs/2004.11867.