Joint Modeling of Chinese Minority Language Translation Tasks

Yifan Guo
Zhengzhou University
Henan, China
570714651@qq.com

Hongying Zan
Zhengzhou University
Henan, China
iehyzan@zzu.edu.cn

Hongfei Xu
Zhengzhou University
Henan, China
0000-0001-8397-1459

***Abstract*—Neural Machine Translation (NMT) normally requires a large amount of parallel corpus to obtain good performance, which is often unavailable for minority languages. Current methods normally pre-train seq2seq models on monolingual data in a denoising manner and then fine-tune the parallel data to improve the performance of low-resource translation. But minority languages used in adjacent areas may co-relate with each other, and jointly modeling them may lead to better performance. In this paper, we propose to improve the performance of Chinese minority language translation with Multilingual NMT (MNMT). As the tokens of the minority languages are not covered by either Chinese BART or mBART and the vocabulary size of the multilingual data exceeds that of the pre-trained model, we map the vocabulary of minority languages to that of the pre-trained BART according to the frequency and enlarge the BART vocabulary by repeating low-frequency tokens respectively to address them. Our experiment results on the CCMT 2023 Chinese minority language translation tasks show that joint modeling can improve the Uyghur-to-Chinese and the Tibetan-to-Chinese tasks by $+2.85$ and $+1.30$ BLEU respectively with BART base, and lead to BLEU scores of 55.48, 53.52, and 48.26 on the Mongolian-to-Chinese, Tibetan-to-Chinese and Uyghur-to-Chinese translation tasks respectively with BART large.**

***Index Terms*—Chinese Bart, Multilingual Neural Machine Translation, Transformer**

I. Introduction

Neural Machine Translation (NMT) models typically rely on much training data to achieve good performance [1], [2]. But for the translation of minority languages, it might be hard to build large scale datasets to well support the training of NMT models. Previous work on improving low-resource translation focuses on either data augmentation [3]–[6], or pre-training [7], [8].

As some minority languages may be frequently used in adjacent areas and affect each other during their evolution, enabling knowledge transfer in their translation may benefit the performance of low-resource languages. In this paper, we propose to improve the performance of Chinese minority language translation by jointly modeling them with a single model in a Multilingual NMT (MNMT) manner [9]–[13].

Corresponding author: Hongfei Xu. This work is partially supported by the National Natural Science Foundation of China (Grant No. 62306284), the Natural Science Foundation of Henan Province (Grant No. 232300421386), and the Henan Provincial Science and Technology Research Project (Grant No. 232102211041).

The tokens of Chinese minority languages are not well covered in the vocabulary of the pre-trained model. We sort the tokens of the Chinese minority languages and the pre-trained model's vocabulary by frequency and map them accordingly when using BART.

We test our approach on the CCMT 2023 Chinese minority language translation tasks (Mongolian (mn) /Uyghur (uy)/ Tibetan (ti)→Chinese Han (zh)) with both multilingual Transformer [14] and Chinese BART [15]. Experiment results show that our approach can lead to significant improvements over the strong BART baseline.

Our main contributions are as follows:

- We propose to improve the performance of minority language translation by joint modeling in an MNMT manner and address the vocabulary mismatching by frequency-based mapping and low-frequency token duplication.
- We test the effectiveness of our method on the CCMT 2023 Chinese minority translation task, and obtain significant BLEU improvements ($+2.85$ and $+1.30$ for uy→zh and the ti→zh respectively) over the strong BART baseline.

II. Related Work

A. Low-resource Translation

[16] fine-tune the model trained on high-resource translation tasks to improve the translation performance on low-resource tasks. BART [15] and T5 [17] pre-train on monolingual data, and [18]–[21] utilize BART for machine translation tasks and obtain significant improvements. [3], [5] use monolingual data through back-translation. [4], [6] obtain pseudo-bilingual parallel data by adding noise to the training data.

B. Multilingual Neural Machine Translation

[22]–[25] show that the bilingual translation model can achieve good translation results. Multilingual models can share language features and representations to improve translation quality for resource-scarce languages [26]–[32]. Most studies focus on how to mitigate this representation bottleneck [9], [33]–[43] with massively MNMT [27]–[32].

[44], [45] show that leveraging the semantic information of high-resource languages can significantly improve the translation performance of low-resource languages. [46] show that leveraging the similarity between related languages provides

a promising approach to addressing low-resource translation. Transferring a pre-trained multilingual NMT model can improve the performance of downstream language pairs [12], [45], [47], especially for low-resource scenarios [48].

III. Joint Training of Minority Language Translation Tasks

MNMT trains a single model [26], [27], [49] on the data of different translation tasks, enabling transfer learning across languages and help improve performance on low-resource tasks. As in our case, the CCMT 2023 Chinese minority language translation tasks all translate into the same language (Chinese Han), the model indeed does not need specific language tokens to indicate the target language, and we can simply concatenate the training of data of different tasks without modifying the model, as shown in Figure 1. But it is also not a problem to translate into more than one language, and we can follow the common practice of replacing the start-of-sentence token with the specific target language token to indicate the translation direction in this case [49].

However, when using the pre-trained BART for the MNMT of minority languages, there are two problems: 1) the vocabulary of the pre-trained model cannot well cover the tokens in the minority languages, and 2) the joint vocabulary of minority languages might be larger than that of the pre-trained model. We address these two issues through frequency-based vocabulary replacement and low-frequency token duplication respectively.

In addition, we also attempted to utilize mbart pretraining models with a wider range of language options in hopes of achieving better performance. Unfortunately, Mbart did not address these issues for the following reasons: 1) The tokenizer used in Mbart does not include vocabulary for these languages. 2) Compared to Chinese BART, mbart needs to allocate its model capacity and capabilities to other languages, resulting in limited ability to handle minority languages and lower performance on Chinese text.

A. Frequency-based Vocabulary Replacement

Most tokens of the minority languages are not in the vocabulary of the pre-trained model. This can be solved by randomly initializing the embeddings of the minority languages' tokens, but using pre-trained embeddings, which are pre-trained with the other model layers rather than random initialization, may lead to better performance.

The problem with using pre-trained embeddings is how to map the embeddings of tokens of the pre-trained model to those of the minority language. Although the languages are different, they are used in adjacent areas describing the same world, and there might be similarities in the word frequency distribution and word co-occurrence distribution between these languages. Based on this assumption, we sort the pre-trained model's tokens and the minority languages' tokens by their frequencies and map the pre-trained model's token to the minority languages' token of the same rank, i.e., high-frequency pre-trained model's tokens are mapped to high-frequency minority languages' tokens and vice versa. We establish a mapping between the indices in the minority dictionary and the Chinese BART's dictionary using word frequency substitution, as show in Figure 2. This approach aims to directly align the minority language dictionary with the Chinese dictionary, thereby leveraging the pre-trained embedddings.

For addressing the issue of word frequency in Chinese, we utilized a 1TB-sized Chinese dataset to perform word frequency counting. The resulting word frequency dictionary was used to indicate the frequency ranking of words in the Chinese BART dictionary. We handled the word frequencies differently for single-language models and multilingual models. 1) Single-language models: We aligned the word frequencies of minority languages with the corresponding entries in the Chinese BART dictionary. 2) Multilingual models: We computed and sorted the weighted word frequencies for various minority languages based on their respective weighted frequencies using Formula 1, where $Freq_{wei}$ represents the weighted frequency of a token, $Freq_{minor}$ represents the actual frequency of the token in the minority language, and SUM_{minor} represents the total number of tokens in that specific minority language.. The weighted frequencies were then aligned with the Chinese BART dictionary.

$$Freq_{wei} = \frac{Freq_{minor}}{SUM_{minor}} \tag{1}$$

An issue with multilingual modeling is that the amounts of data are usually different for different translation tasks, and the task with more training data leads to higher token counts than the others. To address this, we count the tokens independently in each dataset, normalize their counts by the total number of tokens in the dataset, and perform the token replacement based on the normalized token frequency (i.e., the token probability in its corresponding dataset). For a few tokens that appear in more than one dataset, we take the highest token probability among these datasets as its token probability, and averaging is another choice.

B. Vocabulary Enlargement via Low-frequency Token Duplication

To address the issue of larger vocabulary in multilingual models compared to pre-trained models, we duplicate low-frequency tokens of the pre-trained model's vocabulary to expand the vocabulary size. As illustrated in Figure 3, we replicate the word embeddings of the least frequent words in the Chinese dictionary. The number of duplicated tokens is adjusted based on the differences in vocabulary sizes between Chinese and the minority languages. This does not affect the embedding mapping of high-frequency tokens while providing pre-trained embeddings for low-frequency tokens.

IV. Experiment

A. Settings

1) Datasets: We tested the effectiveness of our approach on the CCMT 2023 Chinese minority language

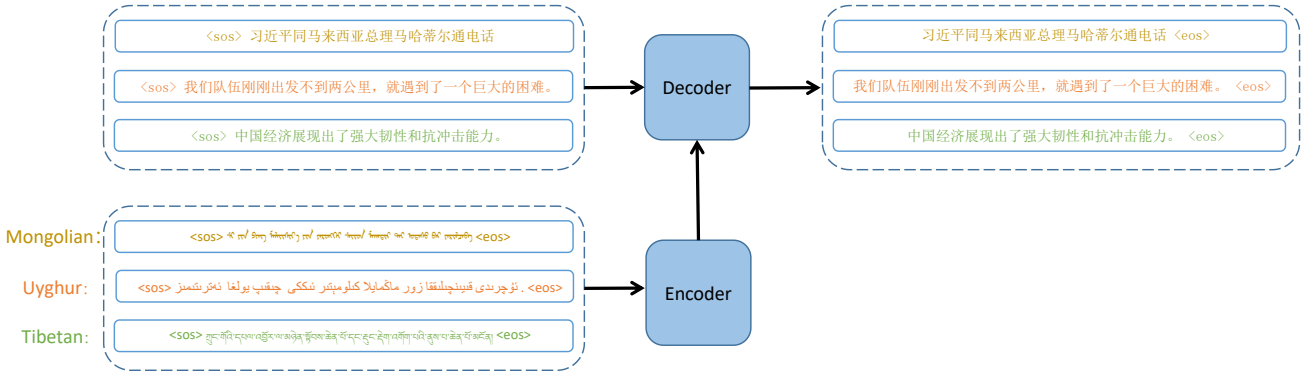

Fig. 1. Multi-language data concatenation.

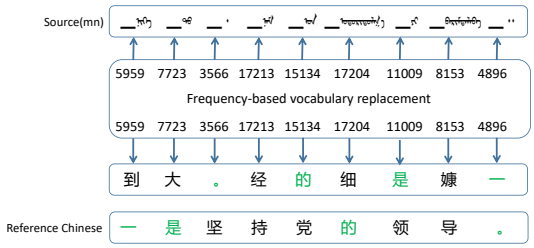

Fig. 2. Frequency-based vocabulary replacement.

TABLE I
STATISTICS OF CCMT 2023 CHINESE MINORITY LANGUAGE TRANSLATION DATASETS.

Task	Training	Validation	Test
mn→zh	1023081	1000	10000
uy→zh	131535	1000	10000
ti→zh	708266	500	10000

TABLE II
MAIN RESULTS.

Methods	mn→zh	uy→zh	ti→zh
Transformer base	53.09	24.75	45.05
+ joint training	53.05	45.19	47.52
+ 24 layers	**56.00**	46.80	48.47
Chinese BART base	51.41	42.96	47.80
+ joint training	51.28	45.81	49.10
+ BART large	55.48	**48.26**	**53.52**

Mongolian→Chinese (mn→zh), Uyghur→Chinese (uy→zh), and Tibetan→Chinese (ti→zh) translation tasks. To address the unknown word issue, we applied independent Byte Pair Encoding (BPE) [50] with the SentencePiece toolkit [51], we used $16k$ merge operations for BPE following [52]. For transformer models, the Chinese Han data were segmented with the jieba toolkit before BPE. For BART models, the Chinese Han data were tokenized by the Chinese BART tokenizer without additional segmentation and BPE.

2) Hyper-parameters: We tested the effectiveness of our approach on both the Transformer model [2] and the Chinese BART model. For Transformer, we used 6 encoder and decoder layers, 512 as the embedding dimension, 4 times of embedding dimension as the number of hidden units of the feed-forward layer, and a dropout probability of 0.1. The number of warm-up steps was set to 8k. We used a batch size of around 25k target tokens achieved by gradient accumulation. For BART, we fine-tuned the model with a fixed learning rate of $1e-5$.

3) Evaluation: We decoded with a beam size of 5 with an average of the last 5 checkpoints saved in an interval of 1500 training steps. As the reference of the test set is not publicly available, we evaluated the translation quality on the validation set by character-level BLEU with the SacreBLEU toolkit [53].

B. Main Results

Our main results are shown in Table II. Table II shows that: 1) our method can lead to consistent and significant improvements on the low-resource uy→zh and ti→zh tasks over both the strong Transformer (+20.44 and +2.47 BLEU respectively) and the BART (+2.85 and +1.30 BLEU respectively) baselines, and the uy→zh task with the least data leads to the largest improvements, 2) On the relatively high-resource mn→zh task, the Transformer outperforms BART fine-tuning, and joint training slightly hampers the performance on the task, while on the low-resource uy→zh and ti→zh tasks, BART consistently outperforms transformer, and

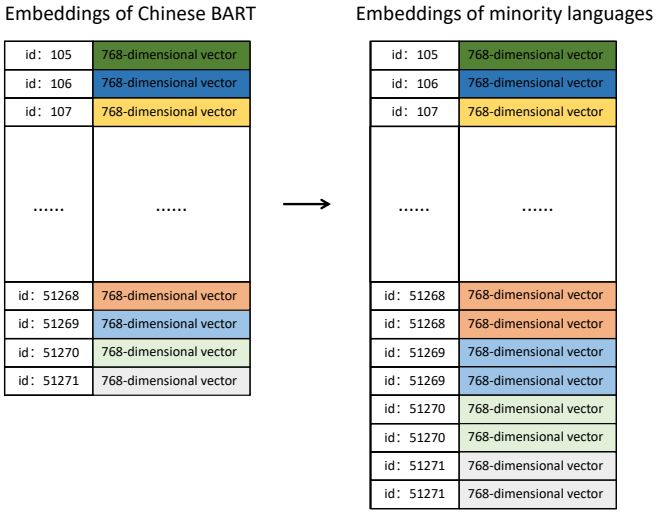

Fig. 3. Vocabulary enlargement.

TABLE III
BLEU RESULTS AFTER DATA REPLACEMENT

Methods	uy→zh	ti→zh
Chinese BART base	42.96	47.80
+ joint training	45.81	49.10
+ replacing mn→zh with en→zh	43.72	46.83

3) using deeper models or larger BART can further improve the performance our approach.

C. Analysis of the Impact of Language

Table II shows that the joint training can improve the performance of low-resource uy→zh and ti→zh tasks using the training data of the mn→zh task. But it is unclear about the impact of the language. We replaced the mn→zh with an equal amount of English→Chinese (en→zh) data extracted from the News Commentary and UN Parallel Corpus of the WMT 2022 news translation task [54]. The en→zh dataset is of a similar domain and high quality. Results are shown in Table III.

Table III depicts that the choice of language has a huge impact on performance. Jointly training on the en→zh data leads to significantly worse performance than on the mn→zh data and even underperforms the BART fine-tuning baseline on the ti→zh task by -0.97 BLEU.

V. Conclusion

We improve the performance of low-resource minority language translation by joint training in an MNMT manner and present frequency-based vocabulary replacement and low-frequency token duplication approaches to addressing the vocabulary coverage and insufficient vocabulary size issues respectively with the pre-trained model for multilingual minority language translation modeling. Our experiments on the CCMT 2023 Chinese minority language translation tasks show that our approach can significantly improve the low-resource uy→zh and ti→zh tasks by $+2.85$ and $+1.30$ BLEU respectively over the strong BART baseline, and find that the choice of the language has a huge impact on the performance.

References

[1] F. Akhbardeh, A. Arkhangorodsky, M. Biesialska, O. Bojar, R. Chatterjee, V. Chaudhary, M. R. Costa-jussa, C. España-Bonet, A. Fan, C. Federmann, M. Freitag, Y. Graham, R. Grundkiewicz, B. Haddow, L. Harter, K. Heafield, C. Homan, M. Huck, K. Amponsah-Kaakyire, J. Kasai, D. Khashabi, K. Knight, T. Kocmi, P. Koehn, N. Lourie, C. Monz, M. Morishita, M. Nagata, A. Nagesh, T. Nakazawa, M. Negri, S. Pal, A. A. Tapo, M. Turchi, V. Vydrin, and M. Zampieri, "Findings of the 2021 conference on machine translation (wmt21)," in *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, November 2021, pp. 1–88. [Online]. Available: https://aclanthology.org/2021.wmt-1.1

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[3] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 489–500. [Online]. Available: https://aclanthology.org/D18-1045

[4] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 567–573. [Online]. Available: https://aclanthology.org/P17-2090

[5] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96. [Online]. Available: https://aclanthology.org/P16-1009

[6] X. Wang, H. Pham, Z. Dai, and G. Neubig, "SwitchOut: an efficient data augmentation algorithm for neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 856–861. [Online]. Available: https://aclanthology.org/D18-1100

[7] M. Neishi, J. Sakuma, S. Tohda, S. Ishiwatari, N. Yoshinaga, and M. Toyoda, "A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size," in *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 99–109. [Online]. Available: https://aclanthology.org/W17-5708

[8] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig, "When and why are pre-trained word embeddings useful for neural machine translation?" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 529–535. [Online]. Available: https://aclanthology.org/N18-2084

[9] B. Zoph and K. Knight, "Multi-source neural translation," *arXiv preprint arXiv:1601.00710*, 2016.

[10] J. Lee, S.-w. Hwang, and T. Kim, "FAD-X: Fusing adapters for cross-lingual transfer to low-resource languages," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online only: Association for Computational Linguistics, Nov. 2022, pp. 57–64. [Online]. Available: https://aclanthology.org/2022.aacl-short.8

[11] C. Baziotis, M. Artetxe, J. Cross, and S. Bhosale, "Multilingual machine translation with hyper-adapters," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1170–1185. [Online]. Available: https://aclanthology.org/2022.emnlp-main.77

[12] S. Sun, A. Fan, J. Cross, V. Chaudhary, C. Tran, P. Koehn, and F. Guzmán, "Alternative input signals ease transfer in multilingual machine translation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5291–5305. [Online]. Available: https://aclanthology.org/2022.acl-long.363

[13] M. Grosso, A. Mathey, P. Ratnamogan, W. Vanhuffel, and M. Fotso, "Robust domain adaptation for pre-trained multilingual neural machine translation models," in *Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 1–11. [Online]. Available: https://aclanthology.org/2022.mmnlu-1.1

[14] B. Zhang, P. Williams, I. Titov, and R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1628–1639. [Online]. Available: https://aclanthology.org/2020.acl-main.148

[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[16] T. Q. Nguyen and D. Chiang, "Transfer learning across low-resource, related languages for neural machine translation," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 296–301. [Online]. Available: https://aclanthology.org/I17-2050

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[18] H. Kim and M. Komachi, "TMU NMT system with Japanese BART for the patent task of WAT 2021," in *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 133–137. [Online]. Available: https://aclanthology.org/2021.wat-1.13

[19] H. Lai, A. Toral, and M. Nissim, "Thank you BART! rewarding pre-trained models improves formality style transfer," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 484–494. [Online]. Available: https://aclanthology.org/2021.acl-short.62

[20] R. Dabre and A. Chakrabarty, "NICT-5's submission to WAT 2021: MBART pre-training and in-domain fine tuning for indic languages," in *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 198–204. [Online]. Available: https://aclanthology.org/2021.wat-1.23

[21] C. Escolano, I. Tsiamas, C. Basta, J. Ferrando, M. R. Costa-jussa, and J. A. R. Fonollosa, "The TALP-UPC participation in WMT21 news translation task: an mBART-based NMT approach," in *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2021, pp. 117–122. [Online]. Available: https://aclanthology.org/2021.wmt-1.6

[22] A. Anastasopoulos, "An analysis of source-side grammatical errors in NMT," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 213–223. [Online]. Available: https://aclanthology.org/W19-4822

[23] J. Zhang and J. van Genabith, "DFKI-NMT submission to the WMT19 news translation task," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 440–444. [Online]. Available: https://aclanthology.org/W19-5350

[24] W. Peng, J. Liu, L. Li, and Q. Liu, "Huawei's NMT systems for the WMT 2019 biomedical translation task," in *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 164–168. [Online]. Available: https://aclanthology.org/W19-5420

[25] R. Rapp, "Similar language translation for Catalan, Portuguese and Spanish using Marian NMT," in *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2021, pp. 292–298. [Online]. Available: https://aclanthology.org/2021.wmt-1.31

[26] O. Firat, K. Cho, and Y. Bengio, "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*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 866–875. [Online]. Available: https://www.aclweb.org/anthology/N16-1101

[27] R. Aharoni, M. Johnson, and O. Firat, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3874–3884. [Online]. Available: https://www.aclweb.org/anthology/N19-1388

[28] B. Zhang, P. Williams, I. Titov, and R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1628–1639. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.148

[29] M. Freitag and O. Firat, "Complete multilingual neural machine translation," in *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 550–560. [Online]. Available: https://www.aclweb.org/anthology/2020.wmt-1.66

[30] K. Ahuja, S. Sitaram, S. Dandapat, and M. Choudhury, "On the calibration of massively multilingual language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4310–4323. [Online]. Available: https://aclanthology.org/2022.emnlp-main.290

[31] A. Mohammadshahi, V. Nikoulina, A. Berard, C. Brun, J. Henderson, and L. Besacier, "SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for

Computational Linguistics, Dec. 2022, pp. 8348–8359. [Online]. Available: https://aclanthology.org/2022.emnlp-main.571

[32] Y. Huang, X. Feng, X. Geng, and B. Qin, "Unifying the convergences in multilingual neural machine translation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6822–6835. [Online]. Available: https://aclanthology.org/2022.emnlp-main.458

[33] G. Blackwood, M. Ballesteros, and T. Ward, "Multilingual neural machine translation with task-specific attention," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3112–3122. [Online]. Available: https://www.aclweb.org/anthology/C18-1263

[34] Y. Wang, J. Zhang, F. Zhai, J. Xu, and C. Zong, "Three strategies to improve one-to-many multilingual translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2955–2960. [Online]. Available: https://www.aclweb.org/anthology/D18-1326

[35] E. A. Platanios, M. Sachan, G. Neubig, and T. Mitchell, "Contextual parameter generation for universal neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 425–435. [Online]. Available: https://www.aclweb.org/anthology/D18-1039

[36] X. Wang, H. Pham, P. Arthur, and G. Neubig, "Multilingual neural machine translation with soft decoupled encoding," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Skeke3C5Fm

[37] X. Tan, Y. Ren, D. He, T. Qin, and T.-Y. Liu, "Multilingual neural machine translation with knowledge distillation," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=S1gUsoR9YX

[38] Y. Wang, L. Zhou, J. Zhang, F. Zhai, J. Xu, and C. Zong, "A compact and language-sensitive multilingual translation method," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1213–1223. [Online]. Available: https://www.aclweb.org/anthology/P19-1117

[39] X. Tan, J. Chen, D. He, Y. Xia, T. Qin, and T.-Y. Liu, "Multilingual neural machine translation with language clustering," 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 963–973. [Online]. Available: https://www.aclweb.org/anthology/D19-1089

[40] A. Bapna and O. Firat, "Simple, scalable adaptation 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1538–1548. [Online]. Available: https://www.aclweb.org/anthology/D19-1165

[41] C. Zhu, H. Yu, S. Cheng, and W. Luo, "Language-aware interlingua for multilingual neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1650–1655. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.150

[42] S. Lyu, B. Son, K. Yang, and J. Bae, "Revisiting Modularized Multilingual NMT to Meet Industrial Demands," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 5905–5918. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.476

[43] H. Xu, Q. Liu, J. van Genabith, and D. Xiong, "Modeling task-aware MIMO cardinality for efficient multilingual neural machine translation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 361–367. [Online]. Available: https://aclanthology.org/2021.acl-short.46

[44] C. Downey, S. Drizin, L. Haroutunian, and S. Thukral, "Multilingual unsupervised sequence segmentation transfers to extremely low-resource languages," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5331–5346. [Online]. Available: https://aclanthology.org/2022.acl-long.366

[45] A. Ebrahimi, M. Mager, A. Oncevay, V. Chaudhary, L. Chiruzzo, A. Fan, J. Ortega, R. Ramos, A. Rios, I. V. Meza Ruiz, G. Giménez-Lugo, E. Mager, G. Neubig, A. Palmer, R. Coto-Solano, T. Vu, and K. Kann, "AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6279–6299. [Online]. Available: https://aclanthology.org/2022.acl-long.435

[46] F. Blum, "Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the low-resource language family tupían," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1–9. [Online]. Available: https://aclanthology.org/2022.acl-srw.1

[47] Y. Kim, Y. Gao, and H. Ney, "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*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1246–1257. [Online]. Available: https://www.aclweb.org/anthology/P19-1120

[48] R. Dabre, A. Fujita, and C. Chu, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1410–1416. [Online]. Available: https://www.aclweb.org/anthology/D19-1146

[49] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017. [Online]. Available: https://www.aclweb.org/anthology/Q17-1024

[50] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[51] T. Kudo and J. Richardson, "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*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[52] W. Hao, H. Xu, L. Mu, and H. Zan, "Optimizing deep transformers for chinese-thai low-resource translation," in *Machine Translation*, T. Xiao and J. Pino, Eds. Singapore: Springer Nature Singapore, 2022, pp. 117–126.

[53] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://aclanthology.org/W18-6319

[54] T. Kocmi, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, T. Gowda, Y. Graham, R. Grundkiewicz, B. Haddow, R. Knowles, P. Koehn, C. Monz, M. Morishita, M. Nagata, T. Nakazawa, M. Novák, M. Popel, M. Popović, and M. Shmatova, "Findings of the 2022 conference on machine translation (wmt22)," in *Proceedings of the Seventh Conference on Machine Translation*. Abu Dhabi: Association for Computational Linguistics, December 2022, pp. 1–45. [Online]. Available: https://aclanthology.org/2022.wmt-1.1