A3-108 Controlling Token Generation in Low Resource Machine Translation Systems

Saumitra Yadav Ananya Mukherjee Manish Shrivastava
MT-NLP Lab
LTRC, KCIS
IIIT Hyderabad, India
saumitra.yadav@research.iiit.ac.in
ananya.mukherjee@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

Translating for languages with limited resources poses a persistent challenge due to the scarcity of high-quality training data. To enhance translation accuracy, we explored controlled generation mechanisms, focusing on the importance of control tokens. In our experiments, while training, we encoded the target sentence length as a control token to the source sentence, treating it as an additional feature for the source sentence. We developed various NMT models using transformer architecture and conducted experiments across 8 language directions (English $\Longleftrightarrow$ Assamese, Manipuri, Khasi, and Mizo), exploring four variations of length encoding mechanisms. Through comparative analysis against the baseline model, we submitted two systems for each language direction. We report our findings for the same in this work.

1 Introduction

Developing Machine Translation solutions for low-resource language pairs is one of the most interesting areas under the umbrella of Machine Translation. There have been many ways of adapting Machine Translation for low-resource language pairs, like,

- Using statistical models instead of neural-based ones to build a system. (Koehn and Knowles, 2017)
- Using multiple combinations of word segmentation to tackle data sparsity in this setting. (Sennrich et al., 2016b; Mujadia and Sharma, 2021; Yadav and Shrivastava, 2021)
- Using monolingual data to create synthetic bitext and train an improved system. (Sennrich et al., 2016a; Burchell et al., 2022; Fadaee et al., 2017)
- Using a pivot language as a bridge between high and low resource language pairs. (Kunchukuttan et al., 2017)
- Using transfer learning (Zoph et al., 2016) by transferring the knowledge from a high language pair setting to a related low language pair setting.
- Multilingual NMT extended on transfer learning by sharing learning space between multiple languages, with the goal of low-resource pair learning from the high-resource pair in a system with decent success. (Johnson et al., 2017)

For low-resource languages, the scarcity of high-quality, extensive datasets necessitates carefully utilising available resources. To maximize the extraction of information from these limited data, we plan to append the target length at the end of the source sentences. This approach draws inspiration from previous research, where incorporating the target length significantly enhanced performance in subtitle generation (Lakew et al., 2019) and current work is adapted from Fan et al. (2018) work on summarization.

In the current work, we consider target token length, length of target sentence **after** subword segmentation, as an additional feature for the source sentence. Intuition is that the system will learn to produce translations subjected to target length. There is an issue of accurately predicting the number of target language tokens in test cases or real-world scenarios. To predict target length, we used multiple methods,

- Neural network to predict target length given source sentence.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 728–734

- Mean token length ratio of target to source sentence from validation set (Lakew et al., 2019) to predict target length given a source sentence.
- Sampling from a normal distribution, where the mean and standard deviation are calculated based on the ratios observed in the validation dataset.
- And for comparison, we also used the actual target length from the test set provided in Pal et al. (2023).

Systems for translating between English and the languages Assamese, Manipuri, Khasi, and Mizo (collectively referred to as IL in the rest of the paper) were developed in this study. It was observed that utilizing the average ratio of target-to-source token lengths from the validation set proved to be an effective method for obtaining control tokens for translation in a low-resource context.

We summarize the contribution of our work as follows.

- Using the number of tokens as a controlling token to improve system performance in a low-resource environment.
- Viable strategy to get control tokens for unseen data.

2 Related Work

Lakew et al. (2019) biased the output length with a transformer architecture using i) target-source length ratio and ii) enriching the transformer positional embedding with length information.

Fan et al. (2018) added the number of tokens to be generated in abstract summarization during training and observed an improvement in the ROUGE score. However, replicating the same for machine translation has been challenging. As Stahlberg (2020) noted, length information can be provided as additional input to the decoder network (Fan et al., 2018; Liu et al., 2018) at each time step as the number of remaining tokens (Kikuchi et al., 2016), or by modifying Transformer positional embeddings (Takase and Okazaki, 2019). Nonetheless, these methods are not directly applicable to machine translation due to the difficulty in accurately predicting translation length.

Additionally, Lakew et al. (2019) biased the output length with a transformer architecture using i) the target-source length ratio and ii) enriching the transformer positional embedding with length information.

3 Approach

This section describes our strategies for computing the control token number, datasets used for training and testing, model architecture, evaluation and systems submitted in shared task.

3.1 Control Tokens

Predicting target length accurately in machine translation remains a complex task, influenced by various factors such as language pair characteristics, sentence structure, and context. To address this challenge, we use a few straightforward heuristics to leverage insights from training and validation data to estimate control tokens effectively. These heuristics aim to give additional information about output length for generation to MT systems. Control tokens (CT) were generated using the following methods (Figure 1):

- **Actual Control Token** refers to the exact count of tokens in the target sentence, derived from a reference or gold standard.
- **Predicted Control Token** is obtained by training a transformer model to predict the number of target tokens given source sentences, where the model learns to estimate the length of the target sentence based on the features extracted from the source sentence. We did this to leverage the self-attention mechanism of the transformer to capture contextual dependencies effectively, making it suitable for tasks requiring an understanding of sentence structure and length prediction.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

- **Ratio Control Token** is the target-to-source token length ratio of the validation dataset for each language pair. Here, we utilize the relationship between the

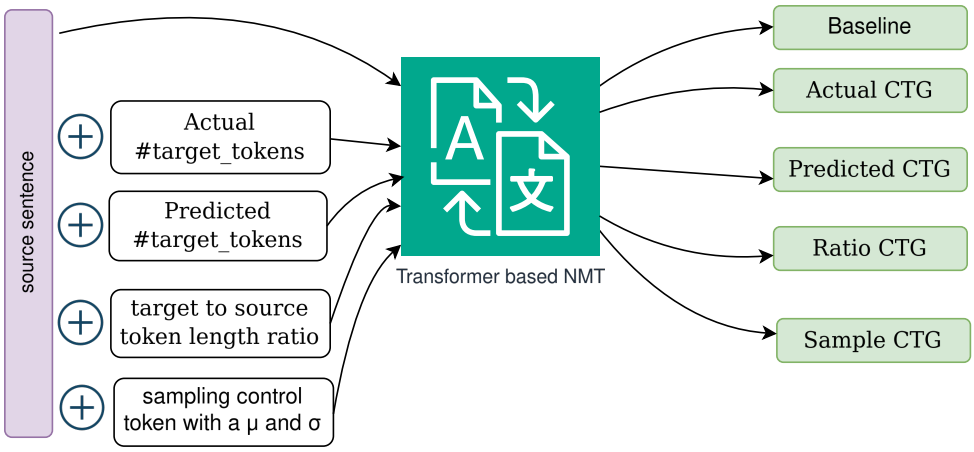

Figure 1: Illustration of our Control Token Generation (CTG) approach

lengths of target sentences and their corresponding source sentences from the validation dataset. For i^{th} source sentence, the control token (CT) is,

$$CT = R_{avg} * len_{source_i} \quad (2)$$

where $R_{avg} = \frac{\sum_{\forall j} len_{target_j}/len_{source_j}}{\text{number of sentence pairs}}$ and, len_{sent} is length of token length of $sent$.

- **Sampled Control Token** is achieved by sampling from a normal distribution where the mean and standard deviation are derived from the ratios observed in the validation dataset.

$$CT = \mathcal{N}(R_{avg}, \sigma^2) * len_{source_i} \quad (3)$$

where $R_{avg} = \frac{\sum_{\forall j} len_{target_j}/len_{source_j}}{\text{number of sentence pairs}}$, σ^2 is standard deviation in ratios and, len_{sent} is token length of sentence $sent$.

3.2 Datasets

We used the Dataset from Pal et al. (2023), Pakray et al. (2024) for English $\Longleftrightarrow$ Assamese, Manipuri, Khasi, and Mizo. Table 1 gives the statistics for each language pair and merge operations (mergeOps) used for the Byte Pair Encoding model of both source and target sentences (Sennrich et al., 2016b).

Figure 2 gives the distribution of IL sentence length with English sentence length ratio for all language pairs. Some sentences in training data have very high ratios compared to validation or test sets. This is where our method can induce learning correspondence between the number of tokens generated and the Control token.

Language Pair	Train	Validation	Test	mergeOps
English Assamese	50 K	2000	2000	16K
English Mizo	50K	2000	1500	16K
English Khasi	24K	1000	1000	4K
English Manipuri	21K	1000	1000	16K

Table 1: Dataset with merge operation for respective language pair

3.3 Architecture

For all the models, we trained machine translation models with the Transformers architecture(Vaswani et al., 2017) using fairseq(Ott et al., 2019) tool[1]. During the training, each source sentence was appended with a 'control token number', the count of target tokens.

4 Experiments

To select the systems as primary and contrastive output, we carried out experiments for English $\Longleftrightarrow$ Assamese, Khasi, Manipuri and Mizo and evaluated the translations of the test set from Pal et al. (2023) using lexical-based metrics, CHRF++ (Popović, 2017).

4.1 Results and Analysis

Table 2 summarises the performance of translation systems for EN-IL and IL-EN using CHRF++. We found statistically significant improvement in translation performance by adding a control token as an additional feature. We observed that,

- In most of the cases, scores improve when the Actual CT is added to the source.

[1]We used basic configuration of transformer architecture

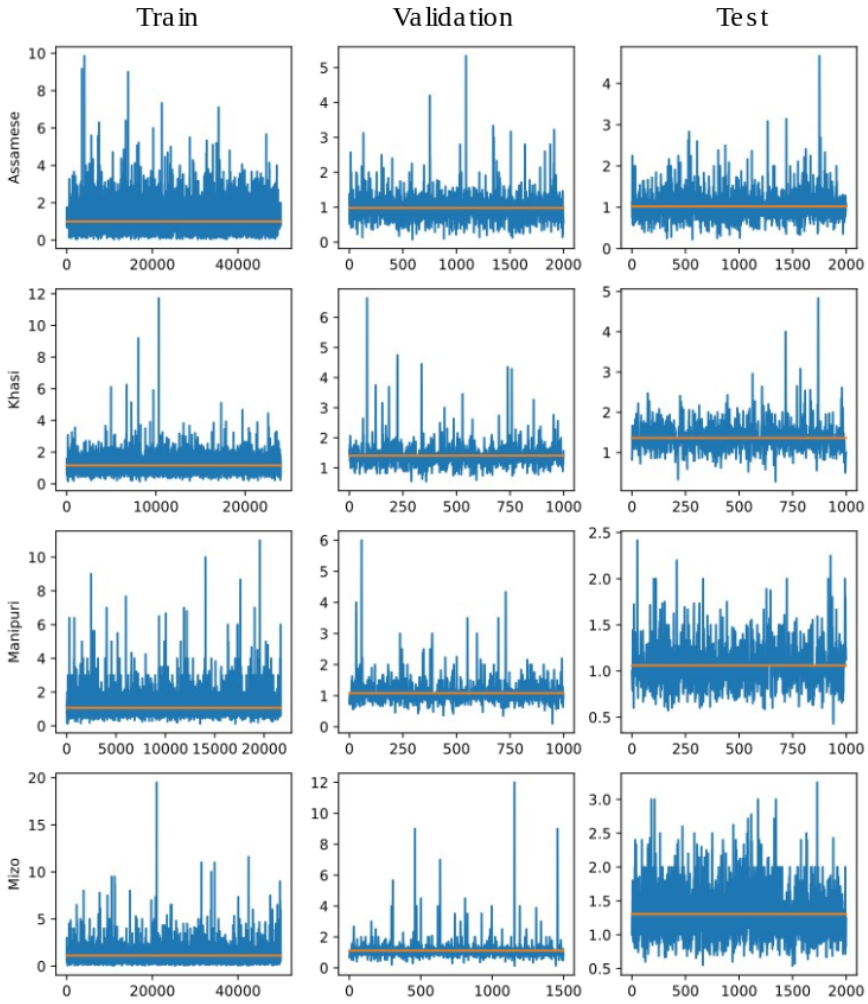

Figure 2: Distribution of Sentence Length Ratio (IL/English) across Train, Validation, and Test Datasets. The orange line denotes the average ratio. The X-axis indicates the number of sentences, and the Y-axis depicts the Sentence Length Ratio (IL/English).

This is expected since it is a gold reference, ensuring the number of target tokens is precise.

- Predicting CT is a challenging problem, as mentioned earlier. While there were improvements in systems, there is also a significant drop in CHRF++ scores for English to Mizo and English to Khasi.

- Utilizing the ratio from validation to determine CT appears to be an optimal choice, which becomes more apparent when examining the distribution of sentence length ratios of Validation and Test in Figure 2. Here, a clear similarity is observed between the two distributions regarding the range of sentence ratios. The inclusion of Ratio CT led to improved performance for English to IL, and vice-versa following actual CT.

- Sampled CT demonstrated strong performance for English to IL, but it did not exhibit the same level of effectiveness for

Language Direction	Baseline	Actual CTG	Predicted CTG	Ratio CTG	Sampling CTG
EN-MZ	**38.11**	37.72	31.44	36.37	37.23
EN-MN	30.47	**32.27**	**31.96**	**32.21**	**32.11**
EN-KH	31.69	**35.35**	26.07	**34.28**	**35.53**
EN-AS	18.44	18.62	18.63	17.67	17.49
MZ-EN	31.03	**32.77**	29.25	**31.5**	**32.68**
MN-EN	**35**	34.24	32.62	34.03	34.45
KH-EN	26.23	**27.58**	26.42	**27.09**	**27.39**
AS-EN	22.76	**24.12**	**23.76**	**23.98**	22.79

Table 2: CHRF++ scores of EN-IL and IL-EN Translation system. Scores in Bold are statistically significant improvements compared to the baseline scores with $p<0.05$.

IL to English. Despite this, it performed comparably well to Ratio CTG in the English to IL direction.

We further analyzed the target-to-source length ratio for EN-IL direction for all 4 language pairs in Figure 3. Examining the average sentence length ratio between Train (●) and baseline systems (■) in comparison to the reference length ratio (▲), sheds light on the behaviour of baseline systems and highlights the advantage of employing CT. In the case of English to Manipuri and Khasi, where significant improvements in CHRF++ scores were noted, the length ratio for baseline systems fell short of the test set. Conversely, when considering the Ratio CTG (◆), we observe their proximity to the Reference Ratio. This supports the idea of using control tokens as an additional feature in the source sentence. It also explains the impact of a poorer prediction system; as seen in the English-to-Mizo ratio, which overshoots by a large margin, there is also a significant drop in translation performance.

Based on these observations, we conclude that if the target sentence length is predictable, leveraging it as an additional feature with the source sentence proves to be a great choice for training a translation model in a low-resource setting.

4.2 Submission

For Translation submission, we preprocessed unseen testset shared by Organizers and submitted translations from the following two systems,

- Primary System: is a model trained using transformer architecture with source sentence and target output length predicted using average **Ratio** of source and target sentences in the validation dataset.
- Contrastive System: is a model trained using transformer architecture without adding CT (Baseline).

Language Direction	**System**	**TER**	**RIBES**	**METEOR**	**ChrF**
English to Assamese	Baseline	100.46	0.0347	0.0587	0.1817
	Ratio	99.79	0.0243	0.05134	0.1773
English to Manipuri	Baseline	101.73	0.0084	0.0179	0.1401
	Ratio	101.55	0.0072	0.0166	0.1415
English to Mizo	Baseline	92.32	0.0406	0.0978	0.18
	Ratio	92.84	0.0328	0.0906	0.173
English to Khasi	Baseline	92.92	0.087	0.1209	0.1905
	Ratio	87.69	0.0873	0.1589	0.2296
Assamese to English	Baseline	96.44	0.0378	0.0677	0.1803
	Ratio	96.19	0.0322	0.0671	0.1883
Manipuri to English	Baseline	96.45	0.029	0.0615	0.1865
	Ratio	96.5	0.0271	0.0635	0.1889
Mizo to English	Baseline	97.75	0.0195	0.0544	0.1633
	Ratio	96.18	0.0181	0.0587	0.1826
Khasi to English	Baseline	105.76	0.0094	0.0403	0.1358
	Ratio	107.7	0.0071	0.0359	0.1348

Table 3: Performance on Unseen Testset

5 Performance on Unseen Testset

Despite the promising results in test sets with training datasets, on the Unseen test set (3) provided by the shared task organizer (Pakray et al., 2024), our approach only gave a slight increase in score compared to the baseline in English to Manipuri, English to Khasi, Assamese to English, Manipuri to English and Mizo to English.

6 Conclusion and Future Work

We address the challenge of translating languages with limited resources by enhancing translation accuracy using target sentence length as an additional feature in the source sentence. We experimented using transformer architecture across 8 language directions (English $\Longleftrightarrow$ Assamese, Manipuri, Khasi, and Mizo). Evaluation against baseline models on a shared test set revealed that our approach significantly improves translation quality in some language directions, demonstrating its effectiveness in improving translation for low-resource languages. However, for the unseen dataset, even though there was an improvement, it wasn't that huge. Overall, we also found that the baseline systems themselves were not promising. Hence, we would be repli-

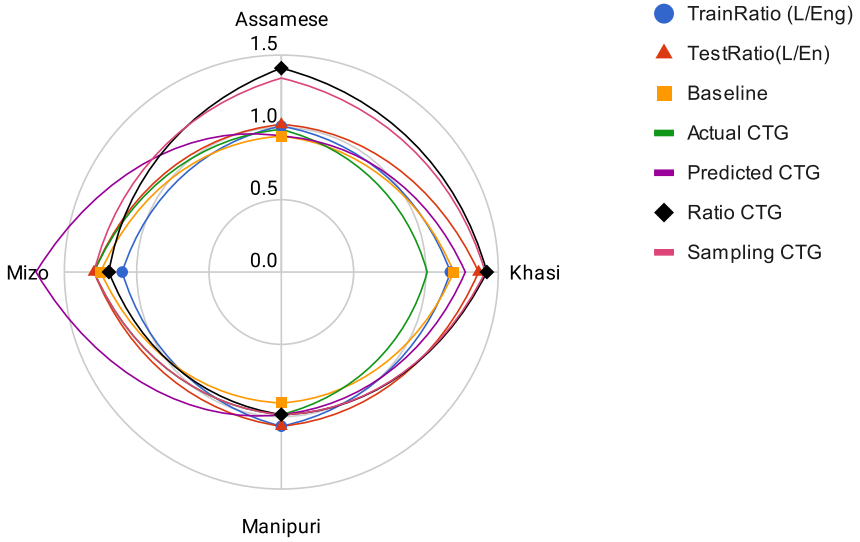

Figure 3: Distribution of Average Sentence length Ratio (IL/English) for Train, Validation and Test Dataset for all language pairs in English to IL direction.

cating this work with other datasets and language pairs to check the validity of this outcome.

References

Laurie Burchell, Alexandra Birch, and Kenneth Heafield. 2022. Exploring diversity in back translation for low-resource machine translation. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 67–79, Hybrid. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. 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)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

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. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

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

Anoop Kunchukuttan, Maulik Shah, Pradyot Prakash, and Pushpak Bhattacharyya. 2017. Utilizing lexical similarity between related, low-resource languages for pivot-based SMT. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 283–289, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. Controlling length in abstractive summarization using a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4110–4119, Brussels, Belgium. Association for Computational Linguistics.

Vandan Mujadia and Dipti Misra Sharma. 2021. English-Marathi neural machine translation for LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 151–157, Virtual. Association for Machine Translation in the Americas.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Mohana Krishna, Arnab Kumar Maji, Sandeep Kumar Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of wmt 2024 shared task on low-resource indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. 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 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Saumitra Yadav and Manish Shrivastava. 2021. A3-108 machine translation system for LoResMT shared task @MT summit 2021 conference. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 124–128, Virtual. Association for Machine Translation in the Americas.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.