MUNI-NLP Systems for Low-resource Indic Machine Translation

Edoardo Signoroni and **Pavel Rychlý**
Faculty of Informatics
Masaryk University
e.signoroni@mail.muni.cz, pary@fi.muni.cz

Abstract

The WMT 2023 Shared Task on Low-Resource Indic Language Translation featured to and from Assamese, Khasi, Manipuri, Mizo on one side and English on the other. We submitted systems supervised neural machine translation systems for each pair and direction and experimented with different configurations and settings for both preprocessing and training. Even if most of them did not reach competitive performance, our experiments uncovered some interesting points for further investigation, namely the relation between dataset and model size, and the impact of the training framework. Moreover, the results of some of our preliminary experiments on the use of word embeddings initialization, backtranslation, and model depth were in contrast with previous work. The final results also show some disagreement in the automated metrics employed in the evaluation.

1 Introduction

This paper describes our systems to the WMT 2023 Shared Task on Low-Resource Indic Language Translation. The task featured four low-resource languages indigenous to the northeastern regions of the Indian subcontinent. The translation was to be done to and from Assamese (Indo-Aryan), Khasi (Austroasiatic), Manipuri, Mizo (Sino-Tibetan) on one side and English on the other. We submitted supervised neural machine translation systems for each pair and direction and experimented with different configurations and settings for both preprocessing and training. We did not use large pretrained models, but trained transformers (Vaswani et al., 2017) of different size and with different parameters for each direction, both on bilingual and multilingual data. Even if most of our final systems did not reach a satisfactory or competitive performance, settling for the middle to low part of the scoreboard, we argue that our experiments brought up some interesting points that call for a deeper investigation. Chiefly, these are the relation between dataset and model size, and the impact of the training framework. Moreover, the final results seem to confirm recent research on the reliability of automatic evaluation metrics, with several cases of disagreements in the ranking of the systems, ranging from one to several places in the leaderboard.

2 Datasets

In the following Section, we first briefly present the languages involved, then we give a summary of the datasets, their contents, domains, and structure.

2.1 Languages

Assamese (*Asamiya*) is an Indo-Aryan language mainly spoken by more than 15 million people in the Indian state of Assam, where it is also the official language. Assamese is also one of the 22 official languages recognized by the Republic of India at the federal level. It is influenced by several other regional languages, mostly Tibeto-Burman varieties, and Bengali, another Indo-Aryan language, with which shares the Bengali-Assamese writing script, an abugida system. Assamese serves as a quasi *lingua franca* for the region and functions as one of the source languages for some pidgins and creoles of the area, such as Nefamese and Nagamese. Assamese is an inflected language with eight grammatical cases and a large collection of classifiers. It follows the subject-object-verb order.

Khasi (*Ka Ktien Khasi*) is an Austroasiatic language with 1 million speakers (the Khasi people) in the Indian state of Meghalaya. It is official in some districts of the state, but not in the state as a whole, and it is considered as "vulnerable". It is related with the other languages in the Khasian group native to the Shillong Plateau, and it is surrounded by unrelated languages such as Assamese, Bengali, Manipuri, and others. It is written both in the Latin, as is the case with this task's data, and the Bengali scripts. Khasi is a stress language

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 959–966
December 6–7, 2023.

Language	ISO-639-3	Family	Script	Num. of Speakers	Official at:	Vitality
Assamese	asm	Indo-Aryan	Bengali	15M	Federal	-
Khasi	kha	Austronesian	Latin	1M	Local	VUL
Mizo	lus	Sino-Tibetan	Latin	1.8M	State	-
Manipuri	mni	Sino-Tibetan	Meitei, Bengali	0.85M	Federal	VUL

Table 1: Summary of the Indic languages involved in the task. For each language, the columns give its ISO code, its language family, the writing system(s) it employs, the number of its speakers and its status, both in terms of official recognition and conservation according to the UNESCO. Khasi and Manipuri are listed as "Vulnerable".

without tones. It has nine grammatical cases and follows the subject-verb-object word order.

Mizo (*Mizo ṭawng*) is a Tibeto-Burman language spoken by around 850 thousand Mizo people, primarily in the Indian state of Mizoram, where it is an official language. It is written in a modified version of the Latin script. Mizo is a tonal language with eight tones, it follows the object-subject-verb order, and it has six grammatical cases.

Manipuri (*Meiteilon*) is a Tibeto-Burman language official in the Manipuri state of India and also at federal level. It is spoken natively by 1.8 million people, the Meitei, both in Manipur and in small communities in the neighboring states. It is considered "vulnerable" by the UNESCO. Manipuri employs a wide array of writing systems, the official ones being the Meitei script and the Bengali script.[1] The Latin script is also used. Manipuri is a tonal language, It follows the subject-object-verb word order.

Table 1 summarizes the main facts about this task's Indic languages.

2.2 Composition

Table 2 gives, for each language pair, the size of the datasets. The parallel datasets made available for this task are small, with the biggest being *asm* and *lus*, at 50k sentence pairs. Of the Indic languages, two are written in Bengali script (*asm* and *mni*), and two in their own variations of the Latin script (*kha* and *lus*). Following the notation of the Flores-200 benchmark dataset (Goyal et al., 2022), we denote the collation of data in Bengali script with *Beng*, and in Latin script with *Latn*.

Monolingual data was released for all Indic languages: *asm*,*lus*, and *mni* have around 2/2.5M sentences each, while *kha* has only 180k. While we did not look at the domains for these data, we sampled the content of the parallel datasets.

[1]This is the writing system used in the task's Manipuri dataset.

Table 3 gives an outline of the contents of each split of each dataset. For *asm*, both the valid and test set differ from the training data. The former is composed mainly by dictionary definitions, while the latter mostly contains religious content. The *kha* dataset is consistent in terms of domains. The *lus* train split has almost exclusively religious content, the validation split contains both religion and instances of single words, and the test split is quite mixed in content. The *mni* data is almost entirely composed by news or otherwise informative text.

3 Methodology

This Section describes our methodology and the baselines we moved from.

3.1 Baselines

For our experiments, we set as our baseline a standard Transformer (Vaswani et al., 2017) with the hyperparameters in Table 4. We wanted to experiment with ways to make the most out of the training data given, and thus we did not use pre-trained models in our work. Almost all models were trained with Fairseq (Ott et al., 2019). The two final submissions to and from Assamese were trained with TorchScale (Ma et al., 2022).

3.2 Preprocessing

The preprocessing for our models was done with SentencePiece (Kudo, 2018), both BPE (Sennrich et al., 2016) and Unigram (Kudo, 2018), and HFT (Signoroni and Rychlý, 2022a). We chose these three segmentation algorithms either for their popularity, as it is the case with BPE and Unigram, or for their stated application, in the case of HFT. For all these algorithms, we set as our baseline parameters a vocabulary size of 2000, with separate dictionaries for source and target language, and a frequency threshold of 100. For other experimental and final runs, we explored different values and settings of segmentation algorithm, vocabulary size and learning, and frequency threshold.

Dataset	Train	Valid	Test	Monolingual	Script
eng-asm	50,000	2,000	2,000	2,624,715	Beng
eng-kha	24,000	1,000	1,000	182,737	Latn
eng-lus	50,000	1,500	2,000	1,909,823	Latn
eng-mni	21,687	1,000	1,000	2,144,897	Beng
eng-Beng	71,687	200	-	-	Beng
eng-Latn	74,000	200	-	-	Latn
eng-all	145,687	400	-	-	Both

Table 2: Size of the dataset for each language pair. Languages are given in ISO-639-3 codes. Train, valid, and test splits are in number of sentence pairs, whereas monolingual data are in number of sentences for the target language. To denote the collation of languages that in the task data are written in Bengali script (asm and mni) and Latin script (kha and lus), we use *Beng* and *Latn*, respectively. *all* denotes the collation of all train splits.

Dataset	Domain		
	Train	Valid	Test
eng-asm	rel,news	misc,news	misc
eng-kha	rel	rel	rel
eng-lus	rel	rel,misc	misc,rel
eng-mni	news	news	news,misc

Table 3: Domains contained in each split of each dataset. Our investigation was conducted on a random sample of each split. *rel(igion)* denotes Bible text and religious news; *news* stands for all non-religious news and information; and *misc* indicates all other miscellaneous domains, e.g. short conversational phrases, dictionary definitions, words.

Parameters	
encoder/decoder layers	6
enc/dec embedding dim	512
enc/dec feed forward dim	2048
enc/dec attention heads	8
optimizer	adam
learning rate	1e-3
warmup updates	4000
dropout	0.3
label smoothing	0.1
max tokens	16384

Table 4: Hyperparameters for our baseline models. Here encoder and decoder parameters are set at the same value.

3.3 Experiments

We explored several ideas and aspects of training during our experiments, which we summarize below.

3.3.1 System Architecture

We tried several configurations of encoder/decoder layers, inspired by previous work such as Araabi and Monz (2020) and van Biljon et al. (2020) which finds that shallower transformers work better in a low-resource scenario. This was the case also for most of our experiments, where smaller models always outperformed the baseline. This holds true even when training on multilingual data. Apart from the baseline, we tested bigger and deeper models, inspired by work such as (Narang et al., 2021; Wei et al., 2022; Wang et al., 2022), on *mni-eng*, which we considered as the "easiest" direction for the models. Preliminary results show degrading performance with the increase of number of encoder layers. We trained on data tokenized with Unigram and a jointly learned vocabulary of 2000, since this was the best performing setup on the validation split. The results of this experiment are given in Table 6, in terms of BLEU score.

One outlier is the translation to and from Assamese, where baseline models, albeit with a dropout of 0.1, outperformed the smaller ones. In these directions, our final systems turned out to be 18/6 models with an embedding dimension of 384 and a feedforward dimension of 1536. However, it should be noted that these final systems were trained in a parallel line of experiments and with a different framework, TorchScale (Ma et al., 2022), which provides further optimization options, such as DeepNorm. Whether the difference in model behavior is due to the difference in training framework is still not clear and could be explored in future work.

3.3.2 Multilingual Training

We trained parent systems on two different multilingual configurations: using all languages in the task, and using only the ones which shared the script. We called these collated dataset *eng-all*, *eng-Beng*, and *eng-Latn* respectively. The intuition here is

	BLEU	ChrF	RIBES	TER	COMET	Place
eng-asm	7.96	27.31	0.31	91.38	0.59	**10/13**
eng-kha	13.90	37.31	0.61	73.99	0.65	**7/11***
eng-lus	20.48	45.60	0.73	61.22	0.68	**9/10**
eng-mni	19.65	53.26	0.66	69.70	0.72	**12/14**
asm-eng	11.29	30.13	0.64	73.39	0.64	**9/13***
kha-eng	12.71	34.55	0.65	78.15	0.56	**6/11***
lus-eng	23.16	43.02	0.72	62.31	0.63	**6/10***
mni-eng	32.18	58.71	0.76	56.35	0.74	**8/14***

Table 5: Summary of the scores of our best submissions reported in the final evaluation. A star (*) denotes the subtasks in which we scored above the organizers' baseline.

enc/dec	BLEU	Increment
4/4	25.89	**+15.41**
6/6	10.48	baseline
8/4	9.52	*-0.96*
12/4	2.95	*-7.53*
16/4	3.08	*-7.4*

Table 6: An example of the effect of changing the depth of the Transformer on the quality of *mni-eng* translation. 4/4 and 6/6 share the same parameters as the final and baseline systems respectively, while other models have an embedding dimension of 384 and a feedforward dimension of 1536.

to leverage script and language relatedness, which we assumed to be present if not for typology, than for script or geographical closeness, in order to obtain better representations of shared subwords and tokens.

We then fine-tuned child systems for each direction, using *eng-Beng* for Assamese and Manipuri, and *eng-Latn* for Khasi and Mizo. *eng-all* was a parent for systems in all directions. We did not specify any language tag or direction for the parent training, since we did not intend to use them for multilingual translation directly. And since the child systems operate only in one direction, we did not need to specify any language tag for fine-tuning either.

Pretraining on all languages proved to be better than standard supervised training for translating into English from Khasi and Mizo, while translation from Manipuri had better performance with the same script parent.

3.3.3 Backtranslation

We experimented with backtranslation in the *eng-mni* direction, by normalizing and deduplicating the provided monolingual data down to around 300k sentence pairs. We then backtranslated the other side with our best available system for the *mni-eng* direction, which had a BLEU score of 32.18. Despite this decent performance, the systems we trained on the backtranslated data, both transformers with 4/4 encoder/decoder layers, embedding dimension of 256 and 384, and feedforward dimension of 1024 and 1536, did not outperform the previous best system. The bigger of the two models had a roughly 2.5 BLEU points on the smaller one, indicating that bigger architectures could have had even better performance. However, we did not test this at this point.

We also tried other back translation approaches, such as Data Diversification (Nguyen et al., 2020), which proved to be effective in the WMT22 Low-resource shared task for Lower/Upper Sorbian and German (Signoroni and Rychlý, 2022b). However, our results using the baseline systems were inconclusive and we decided to explore other approaches.

3.3.4 Word Embeddings Initialization

Previous work (Qi et al., 2018; Edman et al., 2021) showed that using word embeddings to initialize the model's weights improves, sometimes greatly, the performance for low-resource machine translation systems. We tested this in the *eng-mni* direction, training source side word embeddings on the train split with FastText (Bojanowski et al., 2017) in the skipgram setting. While training new baseline systems with the word embedding initialization we observed tiny gains of <0.5 BLEU, however the models converged faster, with 15 to 35 fewer epochs elapsed.

3.3.5 Tokenization Settings

We wanted to explore different settings for the vocabulary size and frequency threshold of the tok-

enizer, as well as for the segmentation algorithm itself, with the objective to find the best settings for each language pair and direction.

Jointly training the vocabulary never resulted in the best system when translating from English, however it gave the best performance for bilingual training of *lus-eng* and *kha-eng*. Nevertheless, these were not the best models overall. With respect to vocabulary size and frequency threshold, in all cases apart from *eng-mni* where we found size 500 and threshold of 200 as best settings, the baselines of 2000 and 100 for these parameters resulted in the best systems. Overall, the picture regarding tokenization and preprocessing settings is not clear and warrants for more investigation.

4 Final Systems

After the experimental phase, we submitted our best performing systems. Table 7 gives their settings and parameters, while Table 5 summarizes the final scores and our placements. Firstly, it should be highlighted that the systems were ranked according to BLEU (Papineni et al., 2002) score. Other metrics, such as ChrF (Popović, 2015), RIBES (Isozaki et al., 2010), TER (Snover et al., 2006), and COMET (Rei et al., 2020), were computed. Looking at the final scores, one can spot several instances in which the metrics do not agree with each other. As a matter of example, the best system for English-Manipuri has 51.96 BLEU, against our twelfth place with 19.65; however our "low-tier" system beats the first one in ChrF (53.26 > 52.61), RIBES (0.66 > 0.51), and COMET (0.72 > 0.57). Recent work has argued for the abandonment of BLEU as a metric of machine translation performance (Kocmi et al., 2021; Mathur et al., 2020; Tan et al., 2015; Sai et al., 2023) in favor of neural metrics which correlate better with human judgements, however this is not always possible when under-resourced languages are involved. While we did not conduct a full and systematic analysis and comparison of the final scores, cases such as the one cited above call for a deeper investigation on automatic evaluation in machine translation.

As our final systems, we obtained roughly two kinds of models: the ones trained on bilingual parallel data, and the ones fine-tuned from a multilingual pair. The former were our best for translating English to the all the Indic languages, and also to translate from Assamese into English. Multilingual pretraining and fine-tuning performed better for the remaining directions, Khasi, Mizo, and Manipuri into English. *kha-eng* and *lus-eng* were fine-tuned from a parent trained on all the parallel dataset, while *mni-eng* was fine-tuned from an Assamese and Manipuri parent. Parent models were trained according to the settings in Table 7, with a patience of 20. fine-tuning was done on only the data for the final translation direction, again with a patience of 20.

Regarding the preprocessing configuration, the settings varied across all the directions. In some cases, such as *eng-kha* and *eng-lus*, sticking to separate source and target vocabulary of size 2000 with a frequency filter of 100 resulted still in the best system. However, for *eng-mni* we found our best system with separate vocabularies of size 500 and a threshold of 200. For multilingual systems, we set the vocabulary size for the Indic side to 750 to try and force the learning of more shared subwords. For the English side, we left the value at 2000. There is no clear winner with respect to segmentation algorithm.

System architecture is the same for all systems, apart from English to and from Assamese. The best architecture was almost always a Transformer with 4 encoder/decoder layers, embedding dimension of 256, feedforward dimension of 1024, and 4 attention heads. For the models involving Assamese, we found that a deeper model of 18 encoder and 6 decoder layers, embedding dimension of 384, feedforward dimension of 1536, and 4 attention heads performed the best.

Other hyperparameters were not investigated extensively, so all of our models were trained with the *adam* optimizer, a learning rate of 1e-3, 4000 warmup updates, a dropout of 0.3, a label smoothing of 0.1, and max tokens for each batch at 16384.

5 Conclusions

This paper describes our experiments and the resulting supervised neural machine translation systems we submitted to WMT23 Low-resource Indic Machine Translation shared task. We trained systems for all directions in the task and experimented with hyperparameter tuning and multilingual training. We did not use transfer learning from pretrained systems, and thus our models were not competitive for some directions. Nonetheless, we argue that our investigation and preliminary analysis on the behavior of different architecture and preprocessing configuration can be useful to other researchers

	eng-asm	eng-kha	eng-lus	eng-mni	asm-eng	kha-eng	lus-eng	mni-eng
training data	bilingual				fine-tuned multilingual			
tokenization	hft			unigram	hft			bpe
src/tgt vocab. size	2000			500	2000	750/2000		
freq. threshold	100			200	100			
enc/dec layers	18/6	4/4			18/6	4/4		
embedding dim.	384	256			384	256		
feedforward dim.	1536	1024			1536	1024		
attention heads	4							
optimizer	adam							
learning rate	1e-3							
warmup updates	4000							
dropout	0.3							
label smoothing	0.1							
max tokens	16384							

Table 7: Summary of the systems for our final submission. The columns give the values for various settings and parameters for preprocessing and training. *bilingual* training denotes a standard supervised training on parallel data, *fine-tuned multilingual* stand for a system fine-tuned on bilingual parallel data, from a parent system trained on more parallel corpora combined. *kha-eng* and *lus-eng* were fine-tuned from a parent trained on all languages, while *mni-eng* parent was trained only on *asm* and *mni* data, which shared the same writing system.

in the field and exposed some interesting points to be explored in future work. Some of our preliminary experiments, such as the use of word embeddings for initialization and backtranslation, did not give the expected results, thus prompting further inquiry.

Limitations and Future Work

As already mentioned above, some instances of disagreement between metrics in the final ranking signal the need for a deeper analysis of the automated evaluation of machine translation. Here, we did not conduct a methodical study on the matter in this instance, this should be the subject for future studies.

The disagreement between metrics notwithstanding, it could be said that overall the performance of our systems was limited. Supervised training showed all its limitations with the small amount of parallel data made available for training. A careful choice of hyperparameters and techniques may ameliorate the situation, but these factors are dependent on the specific dataset involved. Further research must be carried out to uncover clearer connections between the features of the dataset and the choice of parameters and methods to be used. This would cut experimental costs in terms of resources and time, and could lead to better and more efficient models.

However, even if the final systems did not reach competitive levels of performance in some of the cases, our experiments brought up some points that warrant for a deeper investigation. First, the performance of a certain configuration of settings may depend on the framework used for training. The experiments with transformer depth for *mni-eng* contradicts our best systems for Assamese. Whether this discrepancy depends on the languages or on the fact that we used different framework for different translation directions has to be clarified.

Moreover, the connection between dataset and model size has to be investigated further. Assamese worked better with bigger models, even if its dataset was smaller than the multilingual datasets. This goes against the common understanding that a model with fewer parameters is best to deal with fewer data, which will not be enough to train a bigger model. Why this happens only for the Assamese dataset, and not for others, should be better understood.

Ethics Statement

As with any other system trained on real-world data, our models may be biased. These must be taken into account, especially in light of the complex ethnic and religious situation of the region. [2]

Following Lacoste et al. (2019), we report that the experiments and the research that led to the results presented in this paper were conducted

[2]https://www.bbc.com/news/world-asia-india-66086142 (retrieved Aug 31 2023)

on a private server infrastructure consisting of an NVIDIA Tesla T4, A40, and A100 for around 300 hours of training at an efficiency of 0.59 kg/kWh[3] for a total of 44.25 kg CO_2 eq.

References

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lukas Edman, Ahmet Üstün, Antonio Toral, and Gertjan van Noord. 2021. Unsupervised translation of German–Lower Sorbian: Exploring training and novel transfer methods on a low-resource language. In *Proceedings of the Sixth Conference on Machine Translation*, pages 982–988, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

[3]The Czech Republic's country average as reported in https://www.carbonfootprint.com/docs/2018_8_electricity_factors_august_2018_-_online_sources.pdf

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Shuming Ma, Hongyu Wang, Shaohan Huang, Wenhui Wang, Zewen Chi, Li Dong, Alon Benhaim, Barun Patra, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. TorchScale: Transformers at scale. *CoRR*, abs/2211.13184.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, Yanqi Zhou, Wei Li, Nan Ding, Jake Marcus, Adam Roberts, and Colin Raffel. 2021. Do transformer modifications transfer across implementations and applications? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5758–5773, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation.

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.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. 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)*,

pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ananya B. Sai, Vignesh Nagarajan, Tanay Dixit, Raj Dabre, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. Indicmt eval: A dataset to meta-evaluate machine translation metrics for indian languages.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Edoardo Signoroni and Pavel Rychlý. 2022a. HFT: High frequency tokens for low-resource NMT. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 56–63, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychlý. 2022b. MUNI-NLP systems for Lower Sorbian-German and Lower Sorbian-Upper Sorbian machine translation @ WMT22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1111–1116, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Liling Tan, Jon Dehdari, and Josef van Genabith. 2015. An awkward disparity between BLEU / RIBES scores and human judgements in machine translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 74–81, Kyoto, Japan. Workshop on Asian Translation.

Elan van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation.

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.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2022. Deepnet: Scaling transformers to 1,000 layers.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, Jinlong Yang, Miaomiao Ma, Lizhi Lei, Hao Yang, and Ying Qin. 2022. HW-TSC's submissions to the WMT 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 403–410, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.