NLIP_Lab-IITH Low-Resource MT System for WMT24 Indic MT Shared Task

Pramit Sahoo **Maharaj Brahma** **Maunendra Sankar Desarkar**
Natural Language and Information Processing Lab (NLIP)
Indian Institute of Technology Hyderabad
Hyderabad, India
{ai23mtech14004, cs23resch01004}@iith.ac.in, maunendra@cse.iith.ac.in

Abstract

In this paper, we describe our system for the WMT 24 shared task of Low-Resource Indic Language Translation. We consider eng ↔ {as, kha, lus, mni} as participating language pairs. In this shared task, we explore finetuning of a pre-trained machine translation model, where the pretraining objective includes alignment of embeddings of tokens from the 22 scheduled Indian languages by a carefully constructed alignment augmentation strategy (Lin et al., 2020). Our primary system[1] is based on language-specific finetuning on this pre-trained model. We achieve chrF2 scores of 50.6, 42.3, 54.9, and 66.3 on the official public test sets for eng→as, eng→kha, eng→lus, eng→mni respectively. We also explore multilingual training with/without language grouping and freezing of encoder and/or embedding layers.

1 Introduction

The WMT 2024 Shared Task on "Low-Resource Indic Language Translation" (Pakray et al., 2024) extends the efforts in this direction originally initiated in WMT 2023 (Pal et al., 2023), which garnered significant participation from the global community. Recent advancements in machine translation (MT), particularly through techniques like multilingual training and transfer learning, have expanded the scope of MT systems beyond high-resource languages (Johnson et al., 2017). However, low-resource languages continue to present substantial challenges due to the scarcity of parallel data required for effective training (Siddhant et al., 2020; Wang et al., 2022). The shared task focuses on low-resource Indic languages with limited data from diverse language families: Assamese (as), Mizo (lus), Khasi (kha), and Manipuri (mni). The task aims to improve translation quality for the English⇔Assamese, English⇔Mizo, English⇔Khasi, and English⇔Manipuri given the data provided in the constrained setting.

To address the challenges inherent in translating low-resource languages, participants are encouraged to explore several strategies. First, leveraging monolingual data is essential for enhancing translation quality, especially in the absence of sufficient parallel data. Second, multilingual approaches offer the potential for cross-lingual transfer, where knowledge from high-resource languages can be applied to low-resource pairs (Sen et al., 2019). Third, transfer learning provides a mechanism for adapting pre-trained models from high-resource languages to low-resource settings (Wang et al., 2020). Lastly, innovative techniques tailored to low-resource scenarios, such as data augmentation and language-specific fine-tuning, are crucial for improving performance.

In this paper, we describe our system for the WMT 2024 shared task, focusing on fine-tuning two pre-trained models developed by us: IndicRASP and IndicRASP-Seed[2]. IndicRASP model is pre-trained with the objective of aligning embeddings inspired by alignment augmentation (Lin et al., 2020) on 22 Indic languages. Our primary approach involves language-specific fine-tuning, leveraging multilingual training setups, language grouping, and layer freezing. We set up experiments in both bilingual and multilingual settings. We achieve BLEU scores of 20.1 for English→Assamese, 19.1 for English→Khasi, 30.0 for English→Mizo, and 35.6 for English→Manipuri on the public test set, demonstrating the effectiveness of our approach. Specifically, language-specific fine-tuning yielded significant improvements in translation quality, while multilingual setups provided balanced performance across all language pairs. Language grouping and layer freezing are effective techniques

[1] Our code, models, and generated translations are available here: https://github.com/pramitsahoo/WMT2024-LRILT

[2] These pre-trained models were developed for WAT 2024 MultiIndicMT shared task by the authors.

Proceedings of the Ninth Conference on Machine Translation, pages 781–787
November 15-16, 2024

for preserving pre-trained knowledge and mitigating the challenges of multilinguality. Our results highlight the importance of tailored fine-tuning strategies for low-resource languages and show the potential of using alignment-augmented pre-trained models to improve translation quality in low-resource settings.

2 Data

In this section, we present the details of the IndicNECorp1.0 dataset provided by the IndicMT shared task[3] organizers.

Language pair	Script	Dataset	#parallel sents
English-Assamese	Bengali	Training	50000
		Validation	2000
		Test	2000
English-Khasi	Latin	Training	24000
		Validation	1000
		Test	1000
English-Manipuri	Bengali	Training	21687
		Validation	1000
		Test	1000
English-Mizo	Latin	Training	50000
		Validation	1500
		Test	2000

Table 1: Parallel dataset details. Script refers to the writing script of the Indic language.

2.1 Monolingual Data

The official data also includes monolingual data for four languages. The dataset comprises approximately 2.6M sentences for Assamese, 0.1M for Khasi, 2M for Mizo, and 1M for Manipuri.

2.2 Parallel Data

The dataset includes four bilingual pairs between English and Indic languages[4]: English (en) - Assamese (as), English (en) - Khasi (kha), English (en) - Mizo (lus), and English (en) - Manipuri (mni). These languages are mainly spoken in the Northeastern part of India. The English-Assamese and English-Mizo training sets contain 50k parallel sentences each, while the English-Khasi and English-Manipuri training sets contain 24k and 21.6k parallel sentences, respectively. Dataset statistics are presented in Table 1.

3 Approach

In this section, we briefly describe our approaches. We explore transfer learning, language grouping, and layer-freezing techniques.

3.1 Transfer Learning

We explore transfer learning based on two pre-trained models IndicRASP and IndicRASP-Seed. IndicRASP-Seed is a fine-tuned model of IndicRASP on small and high-quality data. Particularly, the pre-trained model is trained on agreement-based objective (Lin et al., 2020; Yang et al., 2020) for Indic languages. Specifically, words from source sentences are randomly substituted by the semantically equivalent words from other languages. The model is pre-trained in 22 scheduled Indic languages using a subset of the Bharat Parallel Corpus Collection (BPCC) dataset (Gala et al., 2023). Out of these 22 languages, two of the shared task languages, Assamese and Manipuri, are part of the pre-training. Alignment augmentation is performed using bi-lingual dictionaries from MUSE[5] (Conneau et al., 2017) and GATITOS[6].

3.2 Language Grouping

We explore the effect of grouping languages based on script similarity in a multilingual setup. Although our primary focus is on bilingual models, for language grouping experiments, we utilize a multilingual approach where languages sharing similar scripts are trained together. This approach is motivated by the idea that joint training with similar languages can improve translation quality due to shared vocabulary and linguistic properties (Jiao et al., 2022; Gala et al., 2023).

- **Group 1** (Bengali script): Assamese and Manipuri
- **Group 2** (Latin script): Khasi and Mizo

3.3 Layer Freezing

We explored layer-freezing approaches to see the impact of freezing different layers of the architecture on final translation performance.

Frozen Encoder: In this approach, we freeze the encoder components during the fine-tuning process to preserve their pre-trained weights from the parent model while the embedding and decoder components are updated.

Frozen Embedding + Encoder: In this setup, we keep the embedding and encoder frozen during

[3]https://www2.statmt.org/wmt24/indic-mt-task.html

[4]Language code as per the dataset provided

[5]https://github.com/facebookresearch/MUSE#ground-truth-bilingual-dictionaries

[6]https://github.com/google-research/url-nlp/tree/main/gatitos

fine-tuning to preserve their pre-trained weights while updating only the parameters of the rest of the layers.

4 Experimental Setup

Settings: We fine-tune pre-trained checkpoints: IndicRASP and IndicRASP-Seed models on official parallel data using the Adam optimizer (Kingma and Ba, 2014) with β_1 set to 0.9 and β_2 set to 0.98. We set the initial warmup learning rate to 1e-07 and the learning rate to 3e-5, with a warmup step of 4000. We train the models with a dropout rate of 0.3 and a label smoothing rate of 0.1. All experiments are conducted on a single NVIDIA A100 GPU. We use a maximum token count of 512 per batch, accumulating gradients over two steps to simulate a larger batch size. The model is trained for up to 1,000,000 updates. We save checkpoints every 2500 updates. We employed a patience of 10 for early stopping.
Evaluation Metrics: We use the official dev and test sets of IndicNECorp1.0 for validation and evaluation. We evaluate using BLEU (Papineni et al., 2002), chrF (Popović, 2015), and chrF++ (Popović, 2017) metrics. We use the SacreBLEU toolkit (Post, 2018) to perform our evaluation[7] with a chrF word order of 2. Additionally, as per the evaluation metrics used by the organizers, we report results on TER (Snover et al., 2006), RIBES (Isozaki et al., 2010), and COMET (Rei et al., 2022) for our primary and contrastive submissions.
Models: We conducted our experiments in both bilingual and multilingual settings. In the bilingual setup, we fine-tuned the IndicTrans2 Distilled model (Gala et al., 2023), IndicRASP, and IndicRASP-Seed models for both English to Indic and Indic to English directions. The translation models are trained separately for each Indic language. In the multilingual setup, we fine-tuned pre-trained checkpoints of IndicRASP and IndicRASP-Seed for both directions. Inspired by Chiang et al. (2022), we initialized the bilingual model with a fine-tuned multilingual model for both English to Indic and Indic to English.

For experiments with layer freezing, we fine-tune pre-trained checkpoints of IndicTrans2 Distilled and IndicRASP-Seed models. Particularly, we perform experiments by freezing the embeddings and encoder and only the encoder component for both English to Indic and Indic to English directions. We conduct all layer-freezing experiments in a bilingual setup. For language grouping experiments, we fine-tune the IndicRASP and IndicRASP-Seed models based on script similarity in a multilingual setup.

5 Results and Discussions

In this section, we report our experimental results and describe our primary and contrastive submissions. The results for our primary and contrastive systems are shown in Table 4. Tables 2, 3, and 5 reports the chrF2, BLEU, and chrF++ scores respectively.

① **English → Indic:** Our primary English to Indic systems are language pair-specific (bilingual models) fine-tuned on pre-trained IndicRASP-Seed, achieving chrF2 scores of 50.6, 42.3, 54.9, and 66.3 for Assamese, Khasi, Mizo, and Manipuri respectively. For the contrastive systems, we consider a bilingual model fine-tuned on a pre-trained IndicRASP checkpoint. The contrastive system achieves chrF2 scores of 49.9, 42.2, 36.5, and 65.8 for Assamese, Khasi, Mizo, and Manipuri, respectively. The detailed primary and contrastive system results are reported in Table 4.

② **Indic → English:** Our primary Indic-to-English systems for Assamese and Manipuri are bilingual models fine-tuned on the pre-trained IndicRASP-Seed model, each achieving chrF2 scores of 52.8 and 67.9, respectively. Similarly, for Khasi and Mizo, our primary systems are bilingual models fine-tuned on a pre-trained IndicRASP checkpoint, achieving a chrF2 score of 36.1 and 49.4, respectively.

For the contrastive Indic-to-English system, we submit a multilingual system fine-tuned on the pre-trained checkpoint of the IndicRASP model, achieving chrF2 scores of 51.2, 36.0, 46.5, and 65.3 for Assamese, Khasi, Mizo, and Manipuri respectively. Table 4 shows the detailed scores in various metrics.

Bilingual vs. Multilingual: We observe IndicRASP-Seed outperforms the IndicRASP model for Assamese and Manipuri. This might be due to the fact that IndicRASP-Seed performs

[7] SacreBLEU signature:
`nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`

Models	English → Indic as	kha	lus	mni	Indic → English as	kha	lus	mni
BILINGUAL SETUP								
INDICTRANS2 DISTILLED FT ON BILINGUAL DATA	49.5	24.9	29.1	60.1	50.9	21.1	22.0	61.9
INDICRASP FT ON BILINGUAL DATA	49.9	42.2	36.5	65.8	50.1	36.1	49.4	67.7
INDICRASP-SEED FT ON BILINGUAL DATA	50.6	42.3	54.9	66.3	52.8	36.1	25.1	67.9
MULTILINGUAL SETUP								
INDICRASP FT ON MULTILINGUAL DATA	49.8	34.6	51.5	63.2	51.2	36.0	46.5	65.3
INDICRASP-SEED FT ON MULTILINGUAL DATA	48.7	34.6	50.2	62.2	52.2	35.3	44.3	65.1
MULTILINGUAL MODEL FT ON BILINGUAL DATA								
INDICRASP MULTILINGUAL MODEL FT ON BILINGUAL DATA	49.3	42.4	54.7	65.8	50.9	36.3	46.8	67.4
LAYER FREEZING								
INDICTRANS2 DISTILLED FT WITH FROZEN ENCODER	47.4	24.4	28.0	57.8	48.7	19.8	18.7	58.8
INDICRASP-SEED FT WITH FROZEN ENCODER	50.4	41.3	48.6	63.4	52.6	26.4	34.2	65.3
INDICTRANS2 DISTILLED FT WITH FROZEN EMBEDDING & ENCODER	46.7	23.1	9.2	15.9	48.8	20.2	19.6	58.1
INDICRASP-SEED FT WITH FROZEN EMBEDDING & ENCODER	50.5	41.2	45.8	62.4	52.9	25.9	29.6	64.1
LANGUAGE GROUPING								
INDICRASP FT WITH SCRIPT SIMILARITY	50.2	35.0	52.1	63.3	52.6	36.4	46.5	66.0
INDICRASP-SEED MODEL FT WITH SCRIPT SIMILARITY	50.3	34.9	53.5	63.6	53.6	36.8	47.4	66.8

Table 2: chrF2 scores on IndicMT WMT24 shared task public test set.

Models	English → Indic as	kha	lus	mni	Indic → English as	kha	lus	mni
BILINGUAL SETUP								
INDICTRANS2 DISTILLED FT ON BILINGUAL DATA	18.0	9.3	13.6	21.6	26.3	2.7	5.0	36.2
INDICRASP FT ON BILINGUAL DATA	20.5	18.9	13.1	33.9	20.0	14.4	29.1	43.6
INDICRASP-SEED FT ON BILINGUAL DATA	20.1	19.1	30.0	35.6	27.4	14.1	6.0	44.1
MULTILINGUAL SETUP								
INDICRASP FT ON MULTILINGUAL DATA	18.7	13.5	25.8	29.0	25.8	14.1	25.4	39.3
INDICRASP-SEED FT ON MULTILINGUAL DATA	17.1	13.2	24.4	27.2	26.7	14.1	23.3	38.3
MULTILINGUAL MODEL FT ON BILINGUAL DATA								
INDICRASP MULTILINGUAL MODEL FT ON BILINGUAL DATA	19.1	19.0	29.7	34.7	25.8	14.8	26.1	43.5
LAYER FREEZING								
INDICTRANS2 DISTILLED FT WITH FROZEN ENCODER	15.6	8.9	13.1	19.6	22.7	1.5	3.0	31.3
INDICRASP-SEED FT WITH FROZEN ENCODER	19.7	18.1	22.4	29.0	26.8	5.6	15.2	40.7
INDICTRANS2 DISTILLED FT WITH FROZEN EMBEDDING & ENCODER	14.8	8.3	2.6	1.3	22.7	1.9	3.8	30.5
INDICRASP-SEED FT WITH FROZEN EMBEDDING & ENCODER	19.4	17.7	19.7	27.2	26.9	5.4	10.9	37.9
LANGUAGE GROUPING								
INDICRASP FT WITH SCRIPT SIMILARITY	19.1	13.8	26.6	28.9	26.9	14.6	25.5	39.8
INDICRASP-SEED MODEL FT WITH SCRIPT SIMILARITY	19.4	14.1	28.6	29.4	28.3	14.8	26.4	40.6

Table 3: BLEU scores on IndicMT WMT24 shared task public test set.

an additional pre-training on a small, high-quality dataset over IndicRASP. However, when the original pre-training dataset did not contain the languages, like the case of Mizo and Khasi languages here, the comparison shows an opposite trend.

Bilingual models perform better than multilingual models, showing a +4.1 and +7.7 chrF2 score improvement for English to Manipuri and English to Khasi, respectively.

Bilingual models initialized with the weights from multilingual models show improvement over the standalone multilingual models, achieving a +7.8 chrF2 score for English to Khasi. This suggests that initializing bilingual models can be helpful in low-resource settings.

Language Grouping: We observe that script-based language grouping shows improvements over a standalone multilingual model with +1.6, +0.3, +3.3, and +1.4 for English to Assamese, Khasi, Mizo, and Manipuri, respectively. It suggests that grouping languages based on script similarity can be effective in addressing the curse of multilinguality.

Layer Freezing: We observe that freezing only the encoder yields better chrF2 scores compared to freezing both the embedding and the encoder. However, layer freezing underperforms compared to full parameter fine-tuned bilingual models.

	BLEU	chrF2	TER	RIBES	COMET
		PRIMARY			
en→as	20.1	50.6	66.0	0.5543	0.8090
en→kha	19.1	42.3	63.5	0.6470	0.6817
en→lus	30.0	54.9	50.0	0.6764	0.7105
en→mni	35.6	66.3	50.5	0.6995	0.7669
as→en	27.4	52.8	65.3	0.6749	0.7854
kha→en	14.4	36.1	82.0	0.5601	0.5773
lus→en	29.1	49.4	66.7	0.6436	0.7004
mni→en	44.1	67.9	50.2	0.7894	0.8162
		CONTRASTIVE			
en→as	20.5	49.9	67.2	0.5356	0.8043
en→kha	18.9	42.2	63.5	0.6499	0.6791
en→lus	13.1	36.5	73.8	0.4357	0.6462
en→mni	33.9	65.8	50.5	0.6972	0.7672
as→en	25.8	51.2	66.8	0.6744	0.7802
lus→en	25.4	46.5	69.0	0.6307	0.6882
mni→en	39.3	65.3	52.4	0.7806	0.8034

Table 4: Submission results on the IndicMT WMT24 public test set.

6 Conclusion

In this paper, we describe NLIP Lab's Indic low-resource machine translation systems for the WMT24 shared task. We explore the translation capabilities of the alignment-augmented pre-trained model, IndicRASP and IndicRASP-Seed, to enhance translation quality for low-resource Indic languages. Experimentally, we found that the IndicRASP model performs better than the IndicTrans2 Distilled model. Additionally, we experiment with layer-freezing and language grouping techniques. In the future, we will focus on refining these techniques and utilizing monolingual data to enhance MT performance for low-resource Indic languages.

Limitations

The pre-trained models use bilingual dictionaries whose domains might differ from the shared task training corpus. Additionally, the considered pre-trained models cover only a limited number of shared task languages. Our submission does not utilize the provided monolingual data, which could further improve model performance through back-translation.

Acknowledgements

We express our gratitude to the reviewer for providing us with valuable feedback and suggestions for improving the readability of the paper. We also thank the Department of Artificial Intelligence and Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, for providing the necessary computing resources to conduct the experiments.

References

Ting-Rui Chiang, Yi-Pei Chen, Yi-Ting Yeh, and Graham Neubig. 2022. Breaking down multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2766–2780, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

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.

Wenxiang Jiao, Zhaopeng Tu, Jiarui Li, Wenxuan Wang, Jen-tse Huang, and Shuming Shi. 2022. Tencent's multilingual machine translation system for WMT22 large-scale African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1049–1056, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. 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

Models	English → Indic as	kha	lus	mni	Indic → English as	kha	lus	mni
Bilingual Setup								
IndicTrans2 Distilled FT on Bilingual data	45.8	25.6	30.3	55.4	49	20	21	59.6
IndicRASP FT on Bilingual data	46.4	41.3	35.2	61.8	46.5	35.3	48.2	65.4
IndicRASP-Seed FT on Bilingual data	47	41.4	53.2	62.3	50.6	35.3	24	65.7
Multilingual Setup								
IndicRASP FT on Multilingual data	46.2	33.4	49.8	58.9	49.1	35.2	45.4	63
IndicRASP-Seed FT on Multilingual data	45.1	33.4	48.5	57.9	50.1	34.6	43.2	62.6
Multilingual Model FT on Bilingual Data								
IndicRASP Multilingual Model FT on Bilingual data	45.7	41.5	53.1	61.9	48.8	35.5	45.7	65.2
Layer Freezing								
IndicTrans2 Distilled FT with Frozen Encoder	43.7	25.1	29.3	53	46.7	18.5	17.6	59.8
IndicRASP-Seed FT with Frozen Encoder	46.8	40.3	46.9	59.1	50.4	25.3	33.1	63
IndicTrans2 Distilled FT with Frozen Encoder & Embeddings	43	24	11.3	13.1	46.8	18.9	18.5	55.6
IndicRASP-Seed FT with Frozen Encoder & Embeddings	46.8	40.2	44.1	58	50.6	24.9	28.6	61.7
Language Grouping								
IndicRASP FT with Script Similarity	46.6	33.8	50.4	59	50.4	35.6	45.4	63.6
IndicRASP-Seed Model FT with Script Similarity	46.7	33.7	51.8	59.4	51.5	36	46.3	64.4

Table 5: chrF2++ scores on IndicMT WMT24 shared task public test set.

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.

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.

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.

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

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. 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.

Wenxuan Wang, Wenxiang Jiao, Shuo Wang, Zhaopeng Tu, and Michael R. Lyu. 2022. Understanding and mitigating the uncertainty in zero-shot translation. *ArXiv*, abs/2205.10068.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online. Association for Computational Linguistics.