OdiaGenAI's Participation at WAT 2025

Debasish Dhal[*,$], Sambit Sekhar[*], Revathy V. R.[#], Akash Kumar Dhaka[†], Shantipriya Parida[†]

[*]Odia Generative AI, Bhubaneswar, India
[$]Aptus Data Labs, Bangalore, India
[#]Department of Computer Science, School of Computational and Physical Sciences, Kristu Jayanti University, Bengaluru, India
[†]AMD Silo AI, Helsinki, Finland

Abstract

This system description paper presents a detailed overview of the model architecture, training procedure, experimental results, and conclusions of the submission from the OdiaGenAI team to the Workshop on Asian Translation (WAT 2025). For this year, we focus only on text-to-text translation tasks for low-resource Indic languages targeting Hindi, Bengali, Malayalam, and Odia languages specifically. The system uses the large language model NLLB-200-3.3B, fine-tuned on large datasets consisting of over 130k rows for each target language. The entire training dataset consists of data provided by the organizers, as in previous years, and augmented by a much larger 100k sentences of data subsampled from the Samanantar dataset provided by AI4Bharat. Our approach achieved competitive BLEU scores on five of the eight evaluation and challenge test submissions.

1 Introduction

Machine Translation (MT) is a long-standing and well-established sub-field within Natural Language Processing dedicated to creating software capable of automatically translating text or speech between languages. Although substantial progress has been made in achieving human-level translation for languages with extensive training corpora, Indic and Asian languages for which much smaller curated corpuses of training data exist still present significant hurdles to existing MT systems and present sufficient scope for improvement (Popel et al., 2020; Costa-jussà et al., 2022). To overcome these challenges and encourage more fruitful research, WAT has served as an open evaluation platform since 2013 (Nakazawa et al., 2020, 2022). While the challenge is multimodal, this year we decided to focus only on the text-to-text translation for the captions present in the dataset ignoring any visual inputs. Just as in the previous yearly submissions, the evaluation of the given translation tasks is conducted using established metrics like Bilingual Evaluation Understudy (BLEU) and Rank-based Intuitive Bilingual Evaluation Scores (RIBES). In this system description paper, we elaborate on our approach to the tasks that we participated in. In comparison to last year, we have added evaluation for Odia while dropping the Hausa language.

- Task 1: English → Hindi (EN-HI) Text only
- Task 2: English → Bengali (EN-BN) Text only
- Task 3: English → Malayalam (EN-ML) Text only
- Task 4: English → Odia (EN-OD) Text only

2 Task Description and Datasets

In addition to the datasets provided by the organizers, for Hindi, Bengali, Odia, and Malayalam, we also used 100k subsampled translation pairs from Samanantar (Ramesh et al., 2022) in the training set, for each of the four languages. As shown in the results section, this was instrumental in improving the results for the fine-tuned models. The training, evaluation and additional challenge splits are detailed in Table 1.

Task 1: English-to-Hindi Translation The organizers provided the HindiVisualGenome 1.1 (Parida et al., 2019)[1] data set (HVG for short). The training part consists of

[1]https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3267

Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025), pages 109–114
December 24, 2025

29k English and Hindi short captions of rectangular areas in photos of various scenes and it is complemented by three test sets: development (D-Test), evaluation (E-Test) and challenge test set (C-Test). Our WAT submissions were for E-Test (denoted "EV" in the official WAT tables) and C-Test (denoted "CH" in the WAT tables).

Task 2: English-to-Bengali Translation For this task, the organizers provided BengaliVisualGenome 1.0 dataset (Parida et al., 2021)[2] (BVG for short). BVG is an extension of the HVG dataset which supports Bengali language. The size of training set and validation set is the same as that for HVG.

Task 3: English-to-Malayalam Translation The organizers provided MalayalamVisualGenome 1.0 dataset[3] (MVG for short). MVG is an extension of the HVG dataset for supporting Malayalam, which belongs to the Dravidian language family (Kumar et al., 2017). The dataset size and images are the same as HVG. MVG contains bilingual English–Malayalam segments, see table 1.

Task 4: English-to-Odia Translation The organizers provided OdiaVisualGenome 1.0 dataset[4] (OVG for short). OVG is a visual genome dataset for Odia language.

3 Modelling and Experimental Details

Identical configurations have been used for all text-to-text translation tasks. For EN-BN, EN-HI, EN-ML, EN-OD text-to-text translation tasks, we individually fine-tuned a large language model (NLLB et al., 2022) separately for all four languages. Similar to Shahid et al. (2023), we used a NLLB-200-3.3B model, but this time chose a much larger 3.3B parameter model, increasing the model size by more than a factor of five. NLLB-200 is a Seq2Seq (Sequence to Sequence) model specifically designed to convert sequences from one domain to sequences in another domain. Bilingual translation (e.g., translating a sequence of words from one language to another) is one of the most prominent applications of Seq2Seq models.

[2]http://hdl.handle.net/11234/1-3722

[3]https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3533

[4]http://hdl.handle.net/11234/1-5979

3.1 Evaluation

As in previous years, the quality of the translation task is evaluated by using the BLEU (Papineni et al., 2002) and RIBES (Wołk and Koržinek, 2016). BLEU is perhaps the most widely used evaluation metric and has been an industry standard for a while. It is widely believed to have good correlation with human evaluation for many language pairs while being fast and easy to compute. RIBES is another popular metric for translation between languages with a different word order where BLEU has been reported to struggle. SacreBLEU is a more recent and standardized variant of BLEU having helped industry with easier reproducibility after a widescale call (Post, 2018).

3.2 Finetuning

Since training all parameters of this large 3.3B model is prohibitively expensive, only a small fraction (0.38%) of the parameters are actually allowed to be tunable while the majority are kept frozen, meaning that their values remain the same during optimization. This is achieved by using LoRA fine-tuning made available through the `peft` package from Huggingface using the `PeftModel` API. All the finetuning runs were executed on 8×AMD Instinct MI250X/MI250 GPUs. Each such GPU unit offers 128GB HBM2e memory with a peak of 362.1 TFLOPS performance using FP16 precision. This computational capacity enabled us to finish each single-language fine-tuning run in approximately eight hours. The hyperparameters used for the fine-tuning runs are presented in Table 4 to facilitate replication.

The training logs for all four runs are presented in figures 1 and 2. The relatively unstable Malayalam-language run (Figure 2) can be attributed to the inherent grammatical complexity of the Dravidian language family. A similar pattern is observed to a smaller extent for the Hindi-language run (Figure 1). We believe that better and higher quality data can improve the performance of the Hindi language. Odia and Bengali-language runs (Figure 2, 1) demonstrate stable training progres-

Set	Sentences	Tokens			
		Bengali	Hindi	Malayalam	Odia
Train (Organizer) (Parida et al., 2019)	28930	113978	145448	107133	141647
Train (Additional) (Ramesh et al., 2022)	100000	1019973	1814937	694570	1025677
Dev	998	3936	4978	3620	4907
Evaluation	1595	6408	7852	5689	7734
Challenge	1400	6657	8639	6044	8100

Table 1: Statistics of our data used in the English→Bengali, English→Hindi, English→Malayalam and English→Odia text-to-text translation task: the number of sentences and tokens.

Language	Visual Genome Source	Samanantar Source	Visual Genome Target	Samanantar Target
Hindi	4.95	16.42	5.03	18.15
Bengali	4.95	11.53	3.94	10.20
Malayalam	4.95	10.19	3.70	6.95
Odia	4.95	11.33	4.90	10.26

Table 2: Average word count for source (English) and target (Indic) sentences across datasets. The word count is calculated by counting the number of words in a sentence, which serves as a proxy for actual token count.

	WAT BLEU		RIBES	
System and WAT Task Label	OdiaGenAI	Best Comp	OdiaGenAI	Best Comp
English→Hindi				
MMEVTEXT21en-hi	45.10	**45.40**	0.831	**0.834**
MMCHTEXT22en-hi	**56.90**	56.10	0.870	**0.870**
English→Bengali				
MMEVTEXT22en-bn	**49.50**	**49.50**	**0.804**	0.801
MMCHTEXT22en-bn	**50.10**	47.50	**0.830**	0.819
English→Malayalam				
MMEVTEXT21en-ml	43.20	**51.20**	0.708	**0.760**
MMCHTEXT22en-ml	**44.20**	40.30	**0.775**	0.757
English→Odia				
MMEVTEXT21en-od	62.90	**64.30**	0.903	**0.906**
MMCHTEXT21en-od	**56.40**	55.40	0.916	**0.916**

Table 3: WAT2025 Automatic and Manual Evaluation Results for English→Hindi, English→Bengali, English→Malayalam and English→Odia text-to-text translation. For each task, we report the scores of our system (OdiaGenAI) alongside those of the best competing submission. The higher score is highlighted in bold. For both metrics, a higher score indicates better performance.

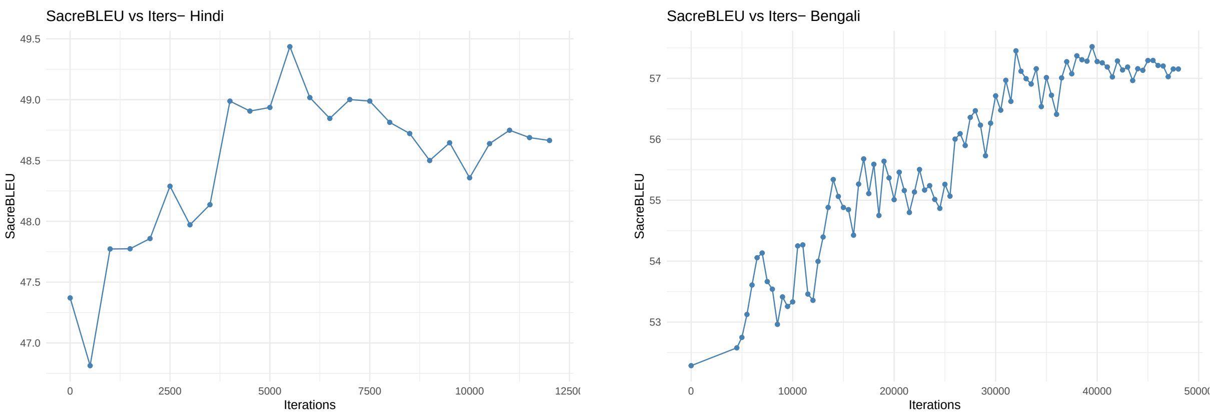

Figure 1: SacreBLEU scores for Hindi and Bengali fine-tuning run.

sion with early convergence, suggesting that extended fine-tuning could yield improved performance. For all four languages, we observe a clear improvement from the starting initial point in the optimization, the highest being for Odia and the lowest for Hindi.

There is still a mismatch in the size of the two components of the final training set. The original dataset provided by the organizers consists of image captions which are short sentences that rarely exceed five words, while the augmented dataset contains many sentences with a higher word count. This case is illustrated in Table 2.

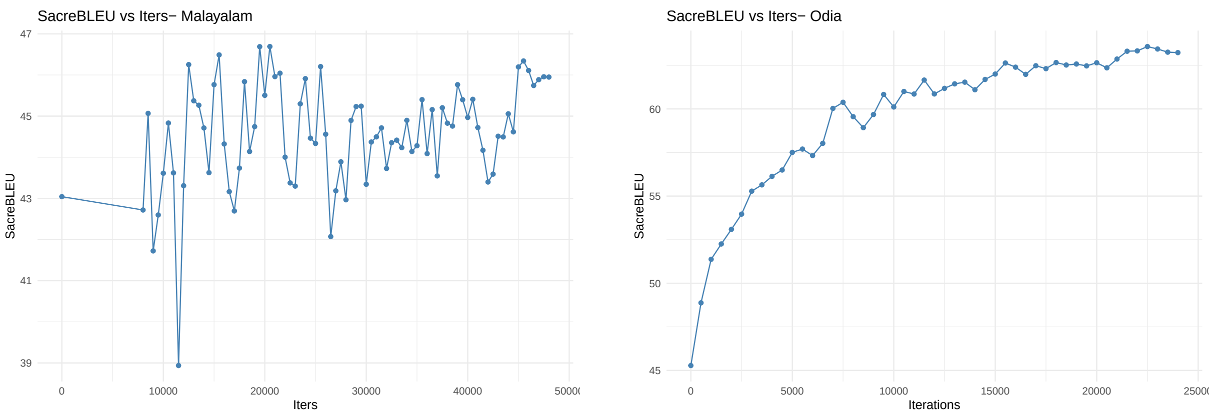

Figure 2: SacreBLEU scores for Malayalam and Odia fine-tuning run.

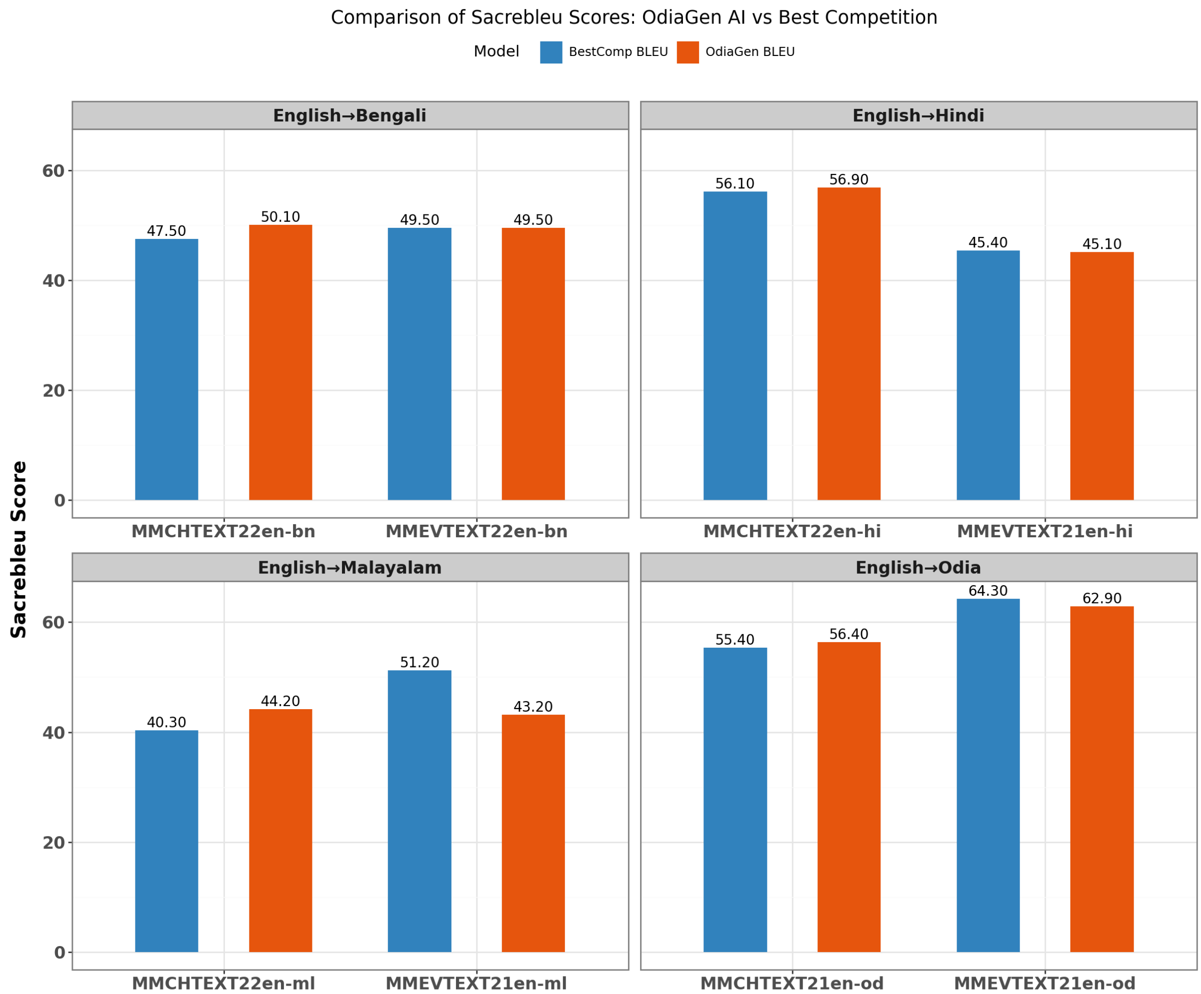

Figure 3: Comparison of our Sacrebleu scores with the best performing team (Source: Table 3).

4 Results

We report the results of the automatic official evaluation after uploading and submitting to the task interface in Table 3, together with the best score attained by the competing submission. Furthermore, we present some selected text samples, translated by our system in Table 5 and do a qualitative analysis. Following the fine-tuning process, these models were used to translate two distinct target test sets for each language: the evaluation set and the challenge set. Translation quality was evaluated using the BLEU score, SacreBLEU,

Hyper Parameter	Value
Learning Rate	$2e^{-4}$
Epochs	3
Cutoff Length	512
Weight Decay	0.01
Warmup Ratio	0.0
max_seq_length	512
LR Scheduler	linear
Lora r	16
Lora α	32
Lora dropout	0.05
use_4bit	False
bnb_4bit_compute_dtype	Not applicable
bnb_4bit_quant_type	None
use_nested_quant	False
per_device_train_batch_size	4 or 8 or 10 or 16
per_device_eval_batch_size	4 or 8 or 10 or 16
gradient_accumulation_steps	1
max_grad_norm	1.0
optim	AdamW
Lora Target Modules	(q_proj, v_proj)

Table 4: Training Hyperparameters.

and RIBES (Ranking by Incremental Bilingual Evaluation System) scores.

For the English-to-Hindi model, a BLEU score of 45.10 was achieved on the evaluation set, while a score of 56.90 was obtained for the challenge set. These results highlight the strong performance of the model and its capacity to handle more complex or unusual translation tasks. The difference between the two scores is 11.8 BLEU points (45.10 vs 56.90) and probably occurs due to a large difference between the two challenge datasets.

In the case of the English-to-Bengali model, a BLEU score of 49.50 and 50.10 were achieved for the evaluation test and challenge sets, respectively. These scores demonstrate strong performance on this task. This indicates a robust overall performance with good generalization and a commendable capability to handle nuanced translations specific to the Bengali language.

BLEU scores of 43.20 and 44.20 were obtained on the evaluation and challenge sets of the Malayalam language, respectively. The best score for the evaluation set of the Malayalam language is 51.20, which is significantly higher than our score.

Our system achieved competitive performance for the Odia language challenge set (56.40), with a BLEU score of 62.90 on the evaluation set. Like the Bengali language, the Odia-language model shows a strong ability for generalized translations.

5 Conclusion

In this system description paper, we presented a system for four text-to-text translation tasks in WAT: (a) English→Hindi, (b) English→Malayalam, and (c) English→Bengali and finally (d) English→Odia text-to-text translation. We released the code through Github for research[5], and the models are released on HuggingFace[6].

These empirical results underscore the effectiveness of the methodology adopted for these MT models. Leveraging a fine-tuned NLLB-200-3.3B model with language-specific Visual Genome datasets provides a robust solution to the MT task for the languages under study: Hindi, Bengali, Malayalam and Odia. The results also pave the way for further enhancements and investigations in the realm of MT.

References

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv e-prints*, pages arXiv–2207.

Arun Kumar, Ryan Cotterell, Lluís Padró, and Antoni Oliver. 2017. Morphological analysis of the Dravidian language family. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 217–222.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation (WAT2022)*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, and 1 others. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44.

[5] https://github.com/shantipriyap/wat2025

[6] https://huggingface.co/collections/OdiaGenAI/wat-2025-finetunedmodels

	Hindi	Bengali	Malayalam	Odia
english-Sentence-1	the orange colored traffic cone	a person wearing a black hat	people on the second level	a water glass on a table
Target-Original	नारंगी रंग यातायात शंकु	একটি কালো টুপি পরা ব্যক্তি	രണ്ടാമത്തെ ലെവലിലെ ആളുകൾ	ଏକ ଟେବୁଲ ଉପରେ ପାଣି ଗ୍ଲାସ ।
Target-Translated	नारंगी रंग का यातायात शंकु	একটি কালো টুপি পরা ব্যক্তি	രണ്ടാം നിലയിലെ ആളുകൾ	ଏକ ଟେବୁଲ ଉପରେ ଏକ ପାଣି ଗ୍ଲାସ ।
Gloss	the orange colored traffic cone	A person wearing a black hat	people on the second level	a water glass on a table
Remarks (Comparison)	Our translation is more grammatically correct	Both are identical	Our translation is fully translated accurately	Our translation is more grammatically correct
english-Sentence-2	the bird is black	This is a person	the court is dark blue	a person walking on a sidewalk
Target-Original	पक्षी काला है	এটি একজন ব্যক্তি	കോർട്ട് ഇരുണ്ട നീല നിറമാണ്	ରାସ୍ତାରେ ଯାଉଥିବା ଜଣେ ବ୍ୟକ୍ତି ।
Target-Translated	पक्षी काला है	এটি একজন ব্যক্তি	കോർട്ട് ഇരുണ്ട നീലയാണ്	ରାସ୍ତାରେ ଯାଉଥିବା ଜଣେ ବ୍ୟକ୍ତି ।
Gloss	the bird is black	This is a person	the court is dark blue	A man walking on the road
Remarks (Comparison)	Both are identical	Both are identical	Both are similar	Both are identical
english-Sentence-3	Man wearing military clothes	A stop light	wooden slat that forms back of bench.	Man wearing military clothes
Target-Original	फौजी कपड़े पहने हुए आदमी	একটি স্টপ লাইট	ഒരു വുഡൻ സ്ലാറ്റ് ബെഞ്ചിന്റെ പുറകിൽ രൂപം കൊള്ളുന്നു.	ସାମରିକ ପୋଷାକ ପିନ୍ଧିଥିବା ବ୍ୟକ୍ତି ।
Target-Translated	सैन्य कपड़े पहने आदमी	একটি স্টপ লাইট	ബെഞ്ചിന്റെ പുറകിൽ രൂപം കൊള്ളുന്ന മരം സ്ലാറ്റ്.	ସାମରିକ ପୋଷାକ ପିନ୍ଧିଥିବା ବ୍ୟକ୍ତି ।
Gloss	Man wearing military clothes	A stop light	Wooden slat that forms the back of the bench.	Man wearing military clothes.
Remarks (Comparison)	Our translation uses a Sanskrit-word for Military, while the target translation uses an Arabic-word.	Both are identical	Our translation is more grammatically correct	Both are identical.

Table 5: Comparison between original translations and our model's translations for English-Malayalam, English-Hindi, English-Bengali, and English-Odia language pairs.

Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multi-modal English to Hindi machine translation. *Computación y Sistemas*, 23(4):1499–1505.

Shantipriya Parida, Subhadarshi Panda, Satya Prakash Biswal, Ketan Kotwal, Arghyadeep Sen, Satya Ranjan Dash, and Petr Motlicek. 2021. Multimodal neural machine translation system for English to Bengali. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 31–39, Online (Virtual Mode). INCOMA Ltd.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

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.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Sk Shahid, Guneet Singh Kohli, Sambit Sekhar, Debasish Dhal, Adit Sharma, Shubhendra Kushwaha, Shantipriya Parida, Stig-Arne Grönroos, and Satya Ranjan Dash. 2023. OdiaGenAI's participation at WAT2023. In *Proceedings of the 10th Workshop on Asian Translation*, pages 46–52, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Krzysztof Wołk and Danijel Koržinek. 2016. Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking. *arXiv preprint arXiv:1601.02789.*