OdiaGenAI's Participation at WAT2023

Sk Shahid Silicon Institute of Technology, Bhubaneswar, India
Guneet Singh Kohli Thapar Institute of Engineering & Technology, India
Sambit Sekhar Odia Generative AI, Bhubaneswar, India
Debasish Dhal NISER, Bhubaneswar, India
Adit Sharma Jaypee Institute of Information Technology, Noida, India
Shubhendra Kushwaha ITER, Siksha 'O'Anusandhan, Bhubaneswar, India
Shantipriya Parida Silo AI, Helsinki, Finland
Stig-Arne Grönroos Silo AI, Helsinki, Finland
Satya Ranjan Dash KIIT University, Bhubaneswar, India

Abstract

This paper offers an in-depth overview of the team "ODIAGEN's" translation system submitted to the Workshop on Asian Translation (WAT2023). Our focus lies in the domain of Indic Multimodal tasks, specifically targeting English to Hindi, English to Malayalam, and English to Bengali translations. The system uses a state-of-the-art Transformer-based architecture, specifically the NLLB-200 model, fine-tuned with language-specific Visual Genome Datasets. With this robust system, we were able to manage both text-to-text and multimodal translations, demonstrating versatility in handling different translation modes.

Our results showcase strong performance across the board, with particularly promising results in the Hindi and Bengali translation tasks. A noteworthy achievement of our system lies in its stellar performance across all text-to-text translation tasks. In the categories of English to Hindi, English to Bengali, and English to Malayalam translations, our system claimed the top positions for both the evaluation and challenge sets.

This system not only advances our understanding of the challenges and nuances of Indic language translation but also opens avenues for future research to enhance translation accuracy and performance.

1 Introduction

Machine translation (MT) is a well-established field within Natural Language Processing (NLP) that focuses on developing computer software to automatically translate text or speech between different languages. While significant progress has been made in achieving human-level translation for high-resource languages, challenges still remain, especially for low-resource languages (Popel et al., 2020; Costa-jussà et al., 2022). Additionally, recent research has explored the effective integration of other modalities, such as images, into the machine translation process.

Proceedings of the 10th Workshop on Asian Translation, pages 46–52
September 4, 2023, Macau SAR, China

The WAT is an open evaluation campaign focusing on Asian languages since 2013 (Nakazawa et al., 2020, 2022). The multimodal translation tasks in WAT2023 consist of image caption translation, in which the input is a descriptive source language caption together with the image it describes, while the output is a target language caption. The multimodal input enables the use of image context to disambiguate source words with multiple senses.

In this system description paper, we (team "ODIAGEN") explains our approach for the tasks (including the sub-tasks) we participated in:

Task 1: English→Hindi (EN-HI) Multimodal Translation

- EN-HI text-only translation
- EN-HI multimodal translation

Task 2: English→Malayalam (EN-ML) Multimodal Translation

- EN-ML text-only translation
- EN-ML multimodal translation

Task 3: English→Bengali (EN-BN) Multimodal Translation

- EN-BN text-only translation
- EN-BN multimodal translation

2 Datasets

We used the datasets specified by the organizer for the related tasks without any additional synthetic data.

Task 1: English→Hindi Multimodal Translation For this task, the organizers provided HindiVisualGenome 1.1 (Parida et al., 2019)[1] dataset (HVG for short). The training part consists of 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 WAT official tables) and C-Test (denoted "CH" in WAT tables).

The statistics of the datasets are shown in Table 1.

Task 2: English→Malayalam Multimodal Translation For this task, the organizers provided MalayalamVisualGenome 1.0 dataset[2] (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. While HVG contains bilingual English–Hindi segments, MVG contains bilingual English–Malayalam segments, with the English, shared across HVG and MVG, see Table 1.

Task 3: English→Bengali Multimodal Translation For this task, the organizers provided BengaliVisualGenome 1.0 dataset[3] (BVG for short). BVG is an extension of the HVG dataset for supporting Bengali. The dataset size and images are the same as HVG, and MVG, see Table 1.

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

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

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

Set	Sentences	Tokens			
		English	Hindi	Malayalam	Bengali
Train	28930	143164	145448	107126	113978
D-Test	998	4922	4978	3619	3936
E-Test	1595	7853	7852	5689	6408
C-Test	1400	8186	8639	6044	6657

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

3 Experimental Details

This section describes the experimental details of the tasks we participated in.

3.1 EN-HI, EN-ML, EN-BN text-only translation

For EN–HI, EN-BN, and EN–ML text-only (E-Test and C-Test) translation, the study fine-tunes the pre-trained NLLB-200 model (NLLB Team et al., 2022), which has been fine-tuned utilizing HVG, BVG, MVG Datasets; aiming to develop a high-quality machine translation system. The NLLB-200 model, a distilled version with 600 million parameters, is used as the base model. It's a Seq2Seq (Sequence-to-Sequence) model, a type of model designed to convert sequences from one domain (like sentences in one language) to sequences in another domain (like sentences in another language). We leverage the Hugging Face's transformers library, specifically using the `AutoModelForSeq2SeqLM` class for the model architecture as shown in Figure 1.

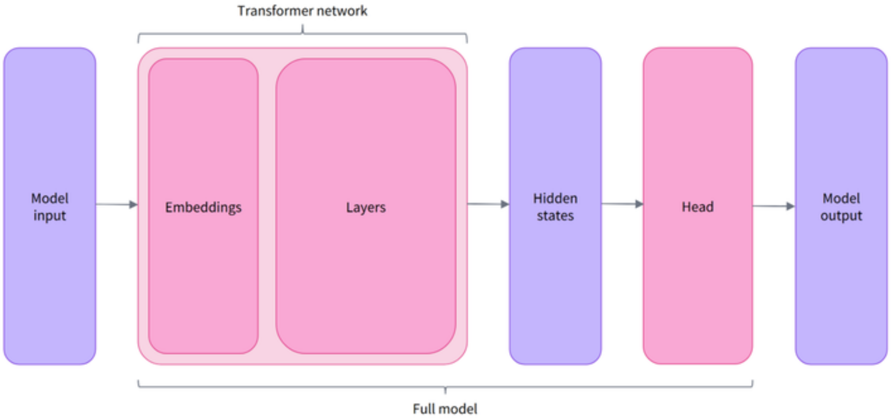

Figure 1: Model Architechture

The pipeline shown in Figure 2 includes several distinct steps:

- **Preprocessing**: The raw text data, which consists of source language sentences and their corresponding target translations, undergoes a preprocessing step. Here, each sentence in both languages is tokenized using a fast tokenizer that leverages **Byte-Pair Encoding (BPE)** (Sennrich et al., 2016). For each sentence, the tokenizer returns an input-ids array, which is a numerical representation of the tokenized sentence, additionally, an attention-mask array is created to indicate the positions of actual tokens. This step results in preprocessed model inputs that include input ids and attention-mask for both source (English) and target (HI/BN/ML) languages.

- **Model Fine-tuning**: The preprocessed inputs are then fed into the NLLB-200 model for training. Given the supervised nature of the task, the model learns to map the source input tokens to the corresponding target tokens. During this process, the model adjusts its internal parameters to minimize the difference between its predictions and the actual target sentences (the labels).

- **Post-processing**: After training, the model generates predictions (preds) for a given English input. These predictions are in the form of token ids, which are then decoded back into their corresponding target sentences using the `tokenizer.batch-decode` function. This decoding process converts the numeric predictions of the model back into human-readable text, ready for evaluation.

- **Evaluation**: Finally, the quality of the model's translations is evaluated using the **Bilingual Evaluation Understudy (BLEU)** score (Papineni et al., 2002). The BLEU score is a popular metric in machine translation that compares machine-generated translations to one or more human-generated reference translations. It provides a quantitative measure of translation quality, with higher scores indicating better performance.

Overall, this pipeline encapsulates the entire process from preprocessing to evaluation, offering a streamlined method for training and validating an English to Hindi/Bengali/Malayalam machine translation model.

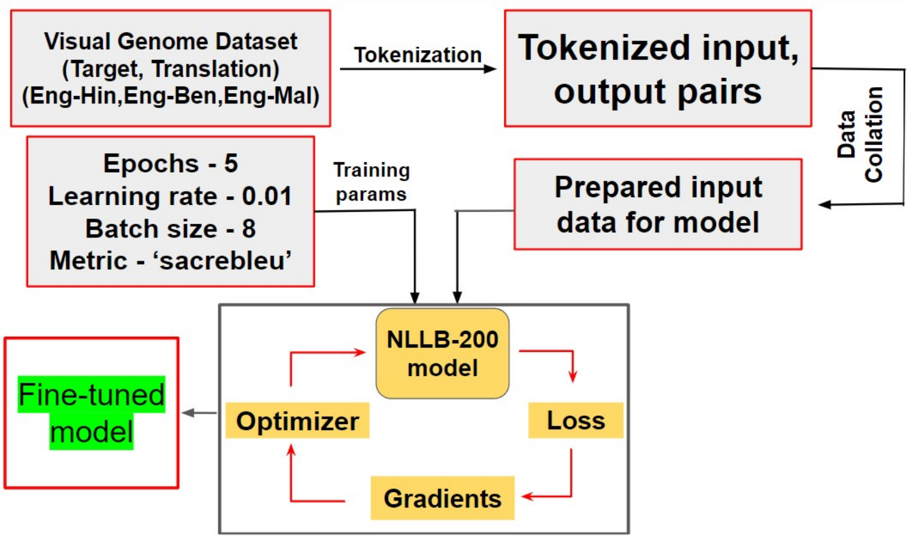

Figure 2: Fine-tuning of NLLB-200 pre-trained model with Visual Genome Dataset

3.2 EN-HI, EN-ML, EN-BN Multimodal translation

This section discusses the multimodal translation pipeline for EN-HI and EN-BN. For EN-HI multimodal (E-Test and CTest) translation, we used the object tags extracted from the HVG dataset images for image features and concatenated them with the text.

Similarly, For EN-BN (E-Test and C-Test) translation, we used object tags extracted from the BVG dataset.

We derive the extracted object tags using a pre-trained Faster RCNN with ResNet101-C4 backbone, which can recognize 80 object types that constitute the COCO Dataset (Lin et al., 2014). In the next step, we select the top 10 tags based on the confidence scores, and in case the object tags are less than 10, we select all the detected tags. The original input English instance is concatenated with a '##' as a separator followed by comma-separated detected tags. This formatted input loaded with visual context from the object tags is fed into the mBART Encoder for processing.

4 Results

We report the official automatic evaluation results of our models for all the participating tasks in Table 2 and sample outputs in Table 3.

Following the fine-tuning process, these models were used to infer translations on two distinct sets for each language: the evaluation set and the challenge set. The translation quality was evaluated using the BLEU (Bilingual Evaluation Understudy) score, and RIBES (Ranking by Incremental Bilingual Evaluation System) score.

For the English-to-Hindi model, a BLEU score of 44.60 was achieved on the evaluation set, while a score of 53.60 was obtained for the challenge set. These results highlight the model's strong performance and its capacity to handle more complex or unusual translation tasks.

In the case of the English-to-Bengali model, a BLEU score of 49.20 was reached on the evaluation set, with a slightly lower score of 47.80 on the challenge set. This indicates a robust overall performance and a commendable capability to handle nuanced translations specific to the Bengali language.

Lastly, for the English-to-Malayalam model, the system achieved a BLEU score of 46.60 on the evaluation set and 39.70 on the challenge set. Despite a slightly lower score on the challenge set, the model still demonstrates a respectable performance in translating English to Malayalam.

Translation Model	**Translation Type**	**BLUE Score (Evaluation Set)**	**BLEU Score (Challenge Set)**
English to Hindi	Text-to-Text	44.60	53.60
	Multimodal	41.60	42.80
English to Bengali	Text-to-Text	49.20	47.80
	Multimodal	42.40	30.50
English to Malayalam	Text-to-Text	46.60	39.70

Table 2: BLEU scores of the text-to-text and multimodal translation models on the evaluation and challenge sets, from the official leaderboard.

The lower BLEU score on the English to Malayalam translation task can be due to a lot of possible factors, one of which is Linguistic Complexity, as Malayalam is a Dravidian language known for its complex grammatical structures and a rich set of linguistic phenomena, which may not be easily captured by the model. This complexity can make the mapping from English to Malayalam challenging.

	MALAYALAM	**HINDI**	**BENGALI**
English-Sentence-1	silver car is parked	fine thin red hair	A stop light
Target-Original	സിൽവർ കാർ പാർക്ക് ചെയ്തു	सूक्ष्म पतले लाल बाल	একটি স্টপ লাইট
Target-Translated	വെള്ളി കാർ പാർക്ക് ചെയ്തിരിക്കുന്നു	ठीक पतले लाल बाल	একটি স্টপ আলো
Gloss	Silver car has been parked	Correct thin red hair	A stop light
Remarks (Comparison)	Translated version is more formal	Original version is better ”Fine” mistranslated by our model.	Original version is more colloquial
English-Sentence-2	eye of the pumpkin	the cross is black	This is a person
Target-Original	മത്തങ്ങയുടെ കണ്ണ്	क्रॉस काला है	এটি একজন ব্যক্তি
Target-Translated	പമ്മിക്കിന്റെ കണ്ണ്	क्रॉस काला है	এটি একজন ব্যক্তি
Gloss	Pumpkin’s eyes	The cross is black	This is a person
Remarks (Comparison)	Model doesn’t translate ”pumpkin”, which is colloquial	Both are identical	Both are identical
English-Sentence-3	pen on the paper	date and time of photo	the bird is black
Target-Original	പേപ്പറിൽ പേന	फोटो की तारीख और समय	পাখিটি কালো
Target-Translated	പേപ്പറിൽ പേന	फोटो की तारीख और समय	পাখিটি কালো
Gloss	Pen on the paper	Date and time of photo	The bird is black
Remarks (Comparison)	Both are identical	Both are identical	Both are identical

Table 3: Comparison between original translations and our model’s translations for English-Malayalam, English-Hindi and English-Bengali language pairs.

5 Conclusion

In this system description paper, we presented our system for three tasks in WAT2023: (a) English→Hindi, (b) English→Malayalam, and (c) English→Bengali Multimodal Translation. We released the code through Github for research[4].

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

Acknowledgements

We are thankful to Silo AI, Helsinki, Finland, and Odia Generative AI, Bhubaneswar, India for the necessary support for participating in WAT2023.

References

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv e-prints*, pages arXiv–2207.

Kumar, A., Cotterell, R., Padró, L., and Oliver, A. (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.

[4] https://github.com/shantipriyap/wat2023

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Nakazawa, T., Mino, H., Goto, I., Dabre, R., Higashiyama, S., Parida, S., Kunchukuttan, A., Morishita, M., Bojar, O., Chu, C., Eriguchi, A., Abe, K., Oda, Y., and Kurohashi, S. (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.

Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., Parida, S., et al. (2020). Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Parida, S., Bojar, O., and Dash, S. R. (2019). Hindi visual genome: A dataset for multimodal English to Hindi machine translation. *Computación y Sistemas*, 23(4):1499–1505.

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., and Žabokrtský, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units.