

Hindi-Odia Machine Translation System

Rakesh Chandra Balabantaray¹, Nayan Ranjan Paul² and Mihir Raj³

¹Computer Science and Engineering, IIT Bhubaneswar, Gothapatna, Bhubaneswar, 751003, Odisha, India

²Computer Science and Engineering, Silicon Institute of Technology, Patia, Bhubaneswar, 751024, Odisha, India

³Computer Science and Engineering, IIT Bhubaneswar, Gothapatna, Bhubaneswar, 751003, Odisha, India

Abstract

This paper provides a comprehensive insight into the translation system developed by team "IIT-BH-MT" for submission to the Machine Translation for Indian Languages(MTIL) track on Forum for Information Retrieval Evaluation (FIRE2023). Our submission is on general translation task and domain specific translation task on Hindi to Odia and Odia to Hindi languages. The system harnesses a cutting-edge Transformer-based architecture, specifically leveraging the NLLB-200 model, which has been fine-tuned using domain-specific and general domain Datasets. The robustness of our system is evident in its proficiency in handling both the translation tasks, demonstrating its versatility across different translation domains. Our results demonstrate robust performance across all evaluated metrics. A standout accomplishment of our system is its exceptional performance across the two translation tasks. Our system secured top positions for both the translation tasks. This system not only help us in understand the difficulties and the intricacies of translation between Hindi-Odia language pairs better, but also paves the way for future research to make translations more accurate and effective.

Keywords

Neural Machine Translation, Fine-tuning, NLLB-200, Machine Translation


1. Introduction


Natural language is a significant part of mankind, covering a wide range of linguistic variations and cultural expression. With more than 23 scheduled languages and 19500 dialects, India, in particular, has an incredibly high level of linguistic variety. The rapid increase in Internet accessibility across the country has enabled people from diverse regions to get extensive access to online services ranging from government services, health care services, etc. However, in order for this accessibility to be genuinely significant, Internet services must be provided in the user's native language, ensuring that the wealth of information can be completely utilized. Translation of important information into local languages has great potential to serve society in areas like agriculture, health, and government among others. Thus, it has become more and more clear that reliable machine translation systems are required, especially those that are made expressly to provide smooth translation across various Indian languages. As we navigate through the nuanced challenges of language diversity in India, our focus is on advancing neural

Forum for Information Retrieval Evaluation, 15-18 December 2023, Panjim, India

✉ rakesh@iit-bh.ac.in (R. C. Balabantaray); nayan.paul@silicon.ac.in (N. R. Paul); mihirraj2119@gmail.com (M. Raj)

ORCID 0000-0002-7421-3805 (N. R. Paul)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

machine translation for Indian language pairs. Machine translation is the need of the hour to bridge the gap between different language pairs.

Machine translation among different Indian languages can be done using multilingual neural machine translation(MNMT). Despite having roots in the NMT period[1, 2], MNMT did not achieve its first major breakthrough until Google's end-to-end MNMT [3]. At that time, the artificial token was first introduced for the translation task at the beginning of the input source sentence to indicate the specified target language, such as "2en" as translating into English. This model employed a common word-piece vocabulary and enabled MNMT via a single encoder and decoder model training. Then via the mBERT-50 [4], M2M-100 [5], and NLLB [6] models. Meta(formerly Facebook) AI expanded the coverage of multilingual translation into 50, 100, and 200+ languages using the later developed structure Transformer and BERT [7, 8]. However, all these models are pre-trained on general domain and not on any specific domain, which sets an obstacle for MT applications in serving to communities on a specific domain.

To encourage researchers for machine translation among Indian languages, FIRE-2023 organized a track named Machine Translation for Indian Languages(MTIL 2023)[9, 10]. The FIRE-2023 MTIL track involves developing a robust machine translation system for translating from the Indian language to the Indian language. There are two tasks in this track. Task 1 is a "General Translation Task" and task 2 is a "Domain Specific Translation Task". Task 1 requires building a machine translation model to translate among 12 Indian language pairs. The task-2 which is domain specific translation task requires to building of machine translation models for the Governance and Healthcare domains. Both the domains have 8 Indian language pairs.

We participated in the FIRE-2023 MTIL for both tasks. In both the tasks our team participated in Odia to Hindi and Hindi to Odia language pairs. Our submissions to the FIRE-2023 MTIL task consist of fine-tuning the NLLB-200[7], a state-of-the-art machine translation model designed specifically for settings with limited resources. NLLB-200 is trained in 1220 language pairs that include 202 languages including Odia and Hindi, which are two of the languages used in the MTIL track. We enhance the model by further training them using data from the general domain. We also train the model on governance domain data and health care domain data. Model assessment is conducted using chrF [11], the official metric for the task. Our model archives the highest chrF for both tasks among all participating groups in Odia to Hindi and Hindi to Odia language pairs.

2. Related Work

This section presents some related literature for machine translation for Hindi-Odia and literature for fine-tuning approaches for MT.

J. Routray et al. [12] presents a shallow parser-based Hindi to Odia MT system. They have used the Apertium platform which is a shallow parser level transformer model. They have also used FST in all modules which makes their the MT model faster. They have also used the TAM(Tense, Aspect, and Modelling) concept in transfer module for building transfer rules between Hindi and Odia in the Apertium platform.

The following literature is available which uses a fine-tuning approach of existing MT models. Gu et al. [13] proposes a model which fine-tuned NLLB-200-600M, a multilingual model for the

task of MT into indigenous American languages. Han et al. [14] proposes transfer learning [15] for NMT via fine-tuning the Meta AI's MPLM model.

By going through the existing literature, there is no work available for machine translation on specific domains particularly for Hindi- Odia language pairs in both directions. This work uses the fine-tuning of the existing NLLB-200 model. This work uses the fine-tuning approach on a dataset of a specific domain to solve tasks of the FIRE-2023 MTIL track.

3. Dataset

Our team has meticulously curated datasets to facilitate training in various domains, focusing on governance, healthcare, and general domains, for both Hindi and Odia language pairs. Within the governance domain, we have created a total of 21,006 sentence pairs. Similarly, for the healthcare domain, our dataset comprises 15,094 sentence pairs. Notably, to encompass a wider scope, we have merged the two sets for the general domain, resulting in a collection of 36,100 sentence pairs. To evaluate the efficiency of our models, we have utilized the dataset provided by the organizers, which includes 1,000 parallel sentences for each domain, ensuring comprehensive testing coverage across all domains.

4. Experimental details

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

The study is dedicated to enhancing the translation capabilities between Hindi and Odia. To achieve this, we have refined the NLLB-200 model, originally developed by the NLLB Team in 2022, through further training. This fine-tuning process involved the utilization of the datasets, with the ultimate goal of creating a notably effective machine translation system for specific domains like governance and healthcare as well as for general domain.

The NLLB-200 model, which serves as the foundation for this study, is a distilled version with a substantial 600 million parameters. It operates on a Seq2Seq (Sequence-to-Sequence) framework, a type of model specifically designed to convert sequences from one domain, such as sentences in one language, to sequences in another domain, such as sentences in another language. The study uses Hugging Face's transformers library, notably leveraging the `AutoModelForSeq2SeqLM` class for the model.

The pipeline shown in Figure 1 includes several distinct steps:

- **Preprocessing:** The raw text data of the dataset undergoes an important preprocessing step to prepare it for the machine translation process. This involves tokenizing each sentence in both the source and target languages using a fast tokenizer based on Byte-Pair Encoding (BPE) [16]. As a result of this tokenization, an array of input-ids is generated for each sentence, representing a numerical form of the tokenized sentence. Additionally, an attention-mask array is created to identify the positions of the actual tokens within the sentences.

In the end, this preprocessing stage produces appropriately processed model inputs that include the input ids and attention mask for the target languages (Odia/Hindi) and the

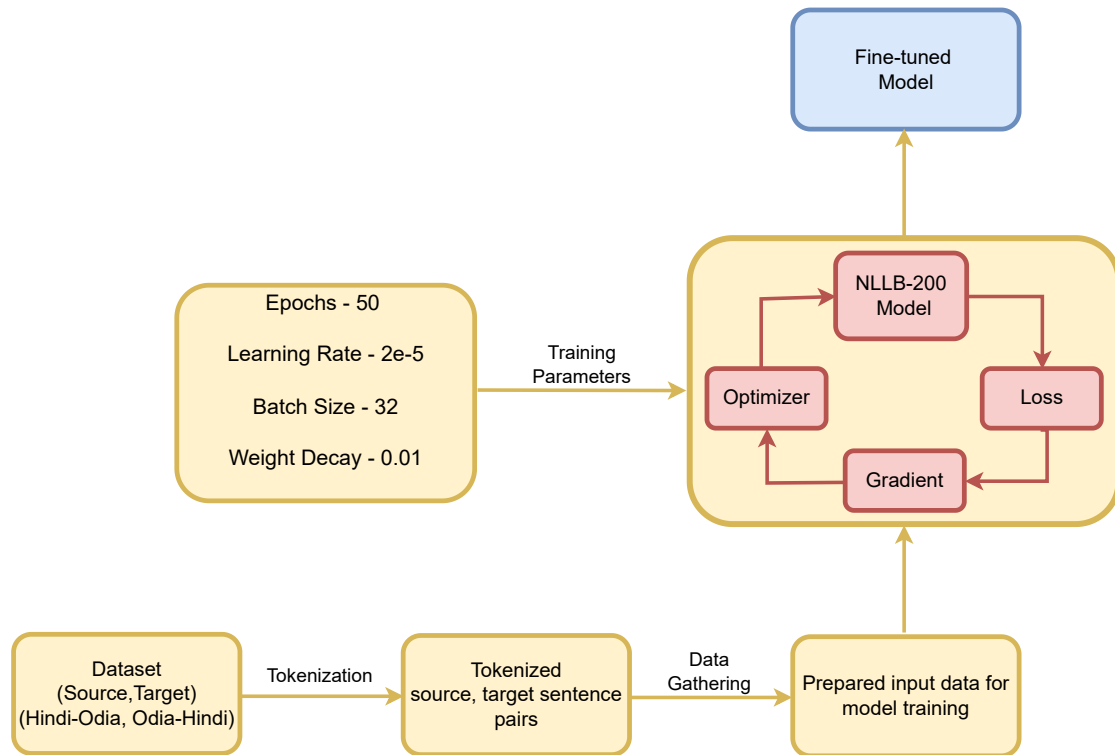


Figure 1: Pipeline for fine-tuning the NLLB-200 Model

source language (Hindi/Odia). These preprocessed inputs make sure the model is ready to handle the translation of text data between the designated languages by laying the groundwork for the next steps of the translation process.

- **Model fine-tuning:** Following the preprocessing stage, the refined inputs are fed into the NLLB-200 model for the purpose of model training. The model performs the complex task of understanding and mapping the source input tokens to their matching target tokens by utilizing a supervised learning framework. During this phase of iterative learning, the model carefully adjusts its internal parameters in order to reduce the difference between the generated predictions and the real target sentences—also referred to as the labels. This thorough optimization not only guarantees the model’s correctness but also develops its ability to distinguish and understand linguistic subtleties, which in turn enables it to provide accurate and dependable translations between the specified source and target languages.
- **Post-processing:** After the training phase, the model creates predictions (called preds) using input from a specific source language (Hndi/Odia). These predictions are originally delivered in the form of token ids, which reflect the model’s generated output. Following that, these token ids are decoded with the help of the "tokenizer.batch-decode" method. Through this important step, the numeric representations of the model’s predictions are

Table 1

Official evaluation scores of two translation tasks on different metrics

Translation Domain	Translation Model	chrF	chrF++	BLEU	TER	COMET
General Domain	Hindi to Odia	61.9976	59.0476	30.5052	51.1916	0.8416
General Domain	Odia to Hindi	66.3945	64.9290	44.1537	41.8758	0.8366
Governance Domain	Hindi to Odia	65.1366	61.9740	32.6069	49.9359	0.8647
Governance Domain	Odia to Hindi	49.2011	48.7519	30.2515	54.2437	0.8576
Healthcare Domain	Hindi to Odia	56.7486	53.5370	23.3734	58.4384	0.8064
Healthcare Domain	Odia to Hindi	60.7855	59.1740	39.2161	49.5659	0.7627

converted back into human-readable text, effectively restoring them to a format that is readily understandable and suitable for detailed evaluation.

- **Model Evaluation:** Finally, the translations produced by the model are assessed using the Bilingual Evaluation Understudy (BLEU) [17], chrF [11], chrF++ [18], COMET [19], and TER scores which are popular machine translation measures. These scores compare machine-generated translations to one or more human-generated reference translations, offering a quantitative assessment of translation quality. Higher scores correspond to better performance.

Overall, this all-inclusive pipeline covers the complete range of tasks, starting with the critical preprocessing phase and ending with the thorough evaluation process. A coherent and systematic strategy for creating a reliable Hindi/Odia to Odia/Hindi machine translation model is shown by combining the data preprocessing, model training, and subsequent validation. This organised approach makes sure that the model is effectively trained and thoroughly tested in order to provide accurate and contextually appropriate translations from Odia to Hindi and from Hindi to Odia.

5. Result Analysis

We present the official evaluation result for all the tasks of our model in table 1.

After the fine-tuning process, these models were employed to generate translations for the test data provided by the FIRE-2023 MTIL track. The quality of the translation was assessed using chrF, BLEU, TER, and COMET. chrF was used as the official metric for evaluation.

In task 1 (General domain), the Hindi to Odia model achieved a chrF score of 61.9976 on the test set, while the Odia to Hindi model scored 66.3945. These results highlight the impressive performance of the models and their capability to handle more complex translation tasks.

In the case of task 2, which comprises two subtasks focused on the Governance and Healthcare domains respectively, the Hindi to Odia model achieved a chrF score of 65.1366 for the Governance domain, while the Odia to Hindi model achieved a score of 49.2011. These results indicate that the Hindi to Odia model performed better for the governance domain compared to the Odia to Hindi model on the respective test set.

In the second subtask focused on the Healthcare domain, the Hindi to Odia model attained a chrF score of 56.7486 on the provided test set, while the Odia to Hindi model achieved a score

of 60.7855. These results of the models exhibit robustness and deliver strong performance on the domain-specific translation task.

6. Conclusion

In this paper, we showcased our entry to the FIRE-2023 MTIL track. Our system leveraged the NLLB-200-600M pre-trained model to perform translations from Hindi to Odia and Odia to Hindi across the General, Governance, and Healthcare domains. Our models demonstrate encouraging outcomes across these tasks, underscoring the effectiveness of the methodology employed for these machine translation models. These empirical results also lay the groundwork for future improvements and exploration in the domain-specific machine translation domain.

Acknowledgments

We thank the Ministry of Electronics and Information Technology (MeitY), Government of India for their assistance through the Project ILMT.

References

- [1] D. Dong, H. Wu, W. He, D. Yu, H. Wang, Multi-task learning for multiple language translation, Association for Computational Linguistics, 2015, pp. 1723–1732. doi:10.3115/v1/P15-1166.
- [2] O. Firat, K. Cho, Y. Bengio, Multi-way, multilingual neural machine translation with a shared attention mechanism, Association for Computational Linguistics, 2016, pp. 866–875. doi:10.18653/v1/N16-1101.
- [3] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean, Google’s multilingual neural machine translation system: Enabling zero-shot translation, Transactions of the Association for Computational Linguistics 5 (2017) 339–351. doi:10.1162/tacl_a_00065.
- [4] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, <https://doi.org/10.48550/arXiv.2008.00401> (2020).
- [5] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, A. Joulin, Beyond english-centric multilingual machine translation, arXiv:2010.11125 (2020).
- [6] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, arXiv:2207.04672v3 [(2022).

- [7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 1706.03762v7 (2017).
- [9] S. Gangopadhyay, G. Epili, P. Majumder, B. Gain, R. Appicharla, A. Ekbal, A. Ahsan, D. Sharma, Overview of mtil track at fire 2023: Machine translation for indian languages, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Goa, India. December 15-18, 2023, ACM, 2023.
- [10] S. Gangopadhyay, G. Epili, P. Majumder, B. Gain, R. Appicharla, A. Ekbal, A. Ahsan, D. Sharma, Overview of mtil track at fire 2023: Machine translation for indian languages, in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, Goa, India. December 15-18, 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.
- [11] M. Popović, chrF: character n-gram f-score for automatic mt evaluation, Association for Computational Linguistics, 2015, pp. 392–395. doi:10.18653/v1/W15-3049.
- [12] J. Rautaray, A. Hota, S. S. Gochhayat, A Shallow Parser-based Hindi to Odia Machine Translation System, 2019, pp. 51–62. doi:10.1007/978-981-10-8055-5_6.
- [13] T. Gu, K. Chen, S. Ouyang, L. Li, Playground low resource machine translation system for the 2023 americasnlp shared task, Association for Computational Linguistics, 2023, pp. 173–176. doi:10.18653/v1/2023.americasnlp-1.19.
- [14] L. Han, G. Erofeev, I. Sorokina, S. Gladkoff, G. Nenadic, Investigating massive multilingual pre-trained machine translation models for clinical domain via transfer learning, arXiv:2210.06068v2 (2022).
- [15] B. Zoph, D. Yuret, J. May, K. Knight, Transfer learning for low-resource neural machine translation, Association for Computational Linguistics, 2016, pp. 1568–1575. doi:10.18653/v1/D16-1163.
- [16] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, Association for Computational Linguistics, 2016, pp. 1715–1725. doi:10.18653/v1/P16-1162.
- [17] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu, Association for Computational Linguistics, 2001, p. 311. doi:10.3115/1073083.1073135.
- [18] M. Popović, chrF++: words helping character n-grams, Association for Computational Linguistics, 2017, pp. 612–618. doi:10.18653/v1/W17-4770.
- [19] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, Comet: A neural framework for mt evaluation, Association for Computational Linguistics, 2020, pp. 2685–2702. doi:10.18653/v1/2020.emnlp-main.213.