CEUR-WS.org/Vol-3681/T9-3.pdf

Fine tuning based Domain Adaptation for Machine Translation of Low Resource Indic Languages

Amulya Ratna Dash, Harpreet Singh Anand and Yashvardhan Sharma

Department of Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani, Jhunjhunu, Rajasthan, India, 333031

Abstract
This paper describes the proposed system for the machine translation of Indic language pairs Odia - Hindi and Hindi - Odia for the General Translation and Domain Specific Translation tasks proposed by Forum of Information Retrieval Evaluation(FIRE) in 2023. For general task, the proposed system uses an ensemble of two pre-trained models and for domain specific task, the proposed system uses a pretrained model fine-tuned using domain specific training data filtered from open source datasets.

Keywords
Low resource Machine Translation, NLLB, BART, IndicTrans, Sentence Similarity

1. Introduction

The importance of language as a mode of communication in the contemporary globalized society cannot be underestimated. In an increasingly globalized and interconnected world, the imperative to overcome linguistic boundaries is paramount in cultivating comprehension, collaboration, and advancement. The Indic languages, are widely spoken by a significant population residing in the Indian subcontinent as well as among diasporic communities. With the fast growing numbers of mobile phone and Internet users, there is an immediate need for automatic machine translation systems from/to English as well as, across Indian languages. Though the digital content in Indian languages has increased a lot in the last few years, it is not yet comparable to that in English. The incorporation of Indic languages into the field of Natural Language Processing (NLP) has been characterized by a gradual and restricted progress, despite the extensive diversity of the Indic linguistic domain.
The 'Machine Translation for Indian Languages' track at FIRE 2023[1][2] consists of two tasks namely General Translation Task(Task 1) and Domain Specific Translation Task(Task 2). Task 1 requires us to build machine translation models to translate sentences of 12 language pairs whereas Task 2 requires us to build machine translation models for Governance and Healthcare domains for 8 language pairs. This paper describes the machine translation system developed for Hindi - Odia and Odia - Hindi language pairs for Task 1 and Task 2.

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

p20200105@pilani.bits-pilani.ac.in (A. R. Dash); f20212416@pilani.bits-pilani.ac.in (H. S. Anand); yash@pilani.bits-pilani.ac.in (Y. Sharma)

© 2021 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)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2. Related Work

The emergence of Encoder-Decoder models, particularly the Transformer neural network architecture proposed by Vaswani et al.[3] in 2017, was a notable breakthrough in the domain of Natural Language Processing (NLP). Attention mechanisms [4] are utilized by transformers in order to process sequences of words in a simultaneous manner, hence enabling the generation of translations that are more contextually relevant and coherent. Transformer-based models have proved to outperform other encoder - decoder models based on RNN and LSTM[5]. The Transformer model uses multi-head self-attention mechanisms and position wise feed-forward networks.

Recent literature indicates that Transformer models pre-trained on large corpora can acquire universal language representations that aid in subsequent tasks. The models are pre-trained on a variety of self-supervised tasks, including predicting a masked word based on its context. Once a model has been pre-trained, it can be fine-tuned on downstream datasets as opposed to being trained from inception. GPT [6][7], BERT[8], and BART[9] are examples of transformer-based pre-trained language models that have had tremendous success in NLP because of their ability to learn universal language representations from large volumes of unlabeled text data and then transfer this knowledge to downstream tasks.

Yin et al.[10] proposed a method for using pre-trained Natural Language Inference(NLI) models as a ready-made zero-shot sequence classifiers. The method works by posing the sequence to be classified as the NLI premise and to construct a hypothesis from each candidate label. IndicBART[11] is a pre-trained BART model for Indic languages, specifically trained for Assamese, Bengali, Gujarati, Hindi, Marathi, Odia, Punjabi, Kannada, Malayalam, Tamil, Telugu and English. Recently transformer-based models specialized for machine translation of Indic languages like IndicTrans[12] and IndicTrans2[13] are available, which are trained on largest available Indic language parallel corpora namely Samanantar and BPCC respectively. IndicTrans model was trained for 11 Indic languages whereas IndicTrans2 was trained fore all the 22 scheduled Indian languages. NLLB(No Language Left Behind)[14], a massively multilingual machine translation model has proven to be a breakthrough in the high-quality translation of around 200 languages across the world. MuRIL(Multilingual Representations for Indian Languages)[15], is a multilingual Language Model specifically built for Indic languages supporting around 17 languages. MuRIL outperforms multilingual BERT (mBERT) on all NLP tasks.

3. Dataset

The dataset utilized for training is extracted from Bharat Parallel Corpus Collection (BPCC)[13], released by AI4Bharat. BPCC is comprised of two parts - BPCC-Mined and BPCC-Human totalling approximately 230 million bi-text pairs. BPCC-Mined contains about 228 million pairs, with nearly 126 million pairs newly added as a part of this work. BPCC-Human, on the other hand consists of 2.2 million gold standard English-Indic pairs, with an additional 644K bitext pairs from English Wikipedia sentences (forming the BPCC-H-Wiki subset) and 139K sentences covering everyday use cases (forming the BPCC-H-Daily subset). However the dataset contains the text in a particular Indian Language and its translation in English. Thus, for training Indic-Indic translation model, we used English as a pivot language to translate Indic-English and then English-Indic language. The availability of direct Indic-Indic parallel dataset may help build a better machine translation model as compared to dataset created via pivoting.

4. Proposed Technique

The proposed technique uses corpus filtering methods, pretrained models, and fine-tuned multilingual models to develop general and domain-specific machine translation systems.

4.1. General Translation Task

The proposed system translates the test set provided by task organizers using NLLB and IndicTrans models. We receive two different versions of translated output for Odia → Hindi and Hindi → Odia using both the models. For Odia → Hindi task, the sentence embeddings of the Odia test set sentences, and their NLLB[1] and IndicTrans[2] Hindi translations using the MuRIL[3] model are generated. Similarly for Hindi → Odia, we generate the sentence embeddings of the Hindi test set sentences, and both versions of Odia translations. Then, we calculate the sentence similarity via embedding cosine similarity of each version of the translations with the corresponding input sentences in the original language and accept the version with higher cross-lingual semantic similarity. We now have Hindi and Odia test data with their most appropriate Odia and Hindi translations respectively.

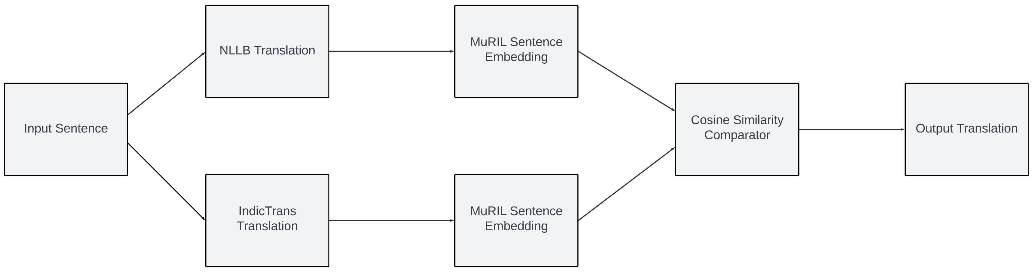

Figure 1: Proposed technique for General Task

4.2. Domain Specific Translation Task

Domain specific task requires translation models specialized for the governance and healthcare domains.

[1] https://huggingface.co/facebook/nllb-200-distilled-600M
[2] https://github.com/AI4Bharat/indicTrans
[3] https://huggingface.co/google/muril-base-cased

4.2.1. Domain specific Dataset

We classified the English sentences from English-Hindi(625K Hindi sentences) and English-Odia(661K Odia sentences) dataset of BPCC using the BART-MNLI[4] model via Zero-Shot Classification.

Table 1
No. of sentences classified for each category

Language	Governance-related	Healthcare-related
Hindi	125587	42445
Odia	109937	74413

The classified sentences in Hindi and Odia are then translated to Odia and Hindi using IndicTrans Model. After the translation the Hindi-Odia and Odia-Hindi synthetic training dataset is split into governance and healthcare specific datasets for fine-tuning the NLLB model.

4.2.2. Fine Tuning of NLLB

The AutoTokenizer from the NLLB model was used to tokenize the inputs. The domain-specific dataset was used to train(fine-tune) the NLLB model in batches of 32 and trained for 5 epochs with a learning rate of 2e-5. Using the same training parameters, we trained four fine-tuned models, Governance specific Hindi-Odia and Odia-Hindi, and Healthcare specific Hindi-Odia and Odia-Hindi model.

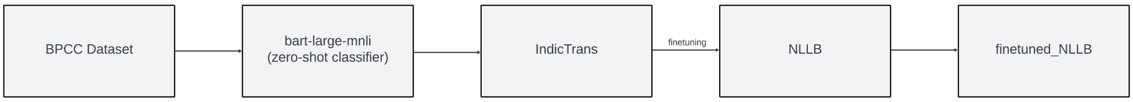

Figure 2: Proposed technique for Domain Specific Task

5. Results

The Table 2 and Table 3 show the official results of our proposed system for General Translation Task and Domain Specific Translation Task respectively.

Table 2
Results for General Task

Model	BLEU	CHRF	CHRF+	TER	COMET
Hindi-Odia	20.057	56.389	51.836	63.967	0.842
Odia-Hindi	29.374	55.572	53.309	56.188	0.804

[4] https://huggingface.co/facebook/bart-large-mnli

Table 3
Results for Domain Specific Task

Model	BLEU	CHRF	CHRF+	TER	COMET
Hindi-Odia (Governance)	23.039	60.327	55.885	61.224	0.867
Odia-Hindi (Governance)	20.031	42.329	40.916	65.476	0.822
Hindi-Odia (Healthcare)	15.225	53.323	48.381	69.468	0.823
Odia-Hindi (Healthcare)	31.931	55.342	53.620	53.791	0.739

6. Conclusion and Future Work

In this paper, we describe our proposed system for machine translation of low-resource Indic language pairs Hindi → Odia and Odia → Hindi, which achieved the second rank (chRF score) for general and domain specific translations in the MTIL track. The proposed system received COMET scores greater than 0.8 on 5 out of 6 sub-tasks, which validates that the translations generated by the models were highly accurate and fluent.
In the future, we would further increase the size of domain-specific training data by exploring other available datasets and data augmentation techniques. Also, we would validate our system for machine translation of other Indic language pairs.

References

[1] 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.

[2] 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.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[4] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, Advances in neural information processing systems 27 (2014).

[5] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[6] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[10] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, arXiv preprint arXiv:1909.00161 (2019).

[11] R. Dabre, H. Shrotriya, A. Kunchukuttan, R. Puduppully, M. M. Khapra, P. Kumar, Indicbart: A pre-trained model for indic natural language generation, arXiv preprint arXiv:2109.02903 (2021).

[12] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. Ak, A. Sharma, S. Sahoo, H. Diddee, D. Kakwani, N. Kumar, et al., Samanantar: The largest publicly available parallel corpora collection for 11 indic languages, Transactions of the Association for Computational Linguistics 10 (2022) 145–162.

[13] J. Gala, P. A. Chitale, R. AK, S. Doddapaneni, V. Gumma, A. Kumar, J. Nawale, A. Sujatha, R. Puduppully, V. Raghavan, et al., Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages, arXiv preprint arXiv:2305.16307 (2023).

[14] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022).

[15] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, et al., Muril: Multilingual representations for indian languages, arXiv preprint arXiv:2103.10730 (2021).