Bridging Language Gaps: Cross-Linguistic Translation Strategies for Indic Languages

Sofia Sunam & Sachin Kumawat & Diksha Seth
Indian Institute of Science
Bengaluru, KA, India
{sofiasunam,sachinkumawa,dikshaseth}@iisc.ac.in

Abstract

Machine Translation is pacing up with other NLP ground-breaking achievements. Although India is a linguistically rich and diverse country, contributing 130 crore to the global speakers, it's high time that we unify the diverse language speakers. AI4Bharat has been recently pouring in efforts to make the base of IndicNLP via various datasets and pre-trained Models. One of the latest cut-through techno model is the Indic Trans2. It is an encoder-decoder Transformer model which has been trained on BPCC, BPCC-BT, etc and supports high-quality translations across all the 22 scheduled Indic languages. Given the need for a unified India, the performance of such a model plays an important role as the model will be utilised for various tasks such as Movie Captioning, Translation of Government policies, Health guidelines, news articles and etc to regional languages. There is a fundamental need for such translations to be robust, accurate and unbiased. There can be a loss of knowledge and waste of resources if any news or career opportunity is not well translated. The misunderstanding and spread of misinformation will create nuances in the country and special care needs to be taken if the cultural sentiments are hurt by such erroneous translations. Not to mention, India has a specific set of English usage which is different and at times invalid to Native English. We tend to improve upon some of the aspects of translation that we found erroneous in the paper Gala et al. (2023).

1 Introduction

India is a linguistically diverse country with numerous local languages. Despite this diversity, these local languages are low-resourcedAdelani et al. (2022) and there are limited resources available for machine translation among these languages. This project aimed to address this gap by developing a better machine translation model that translates Hindi to other local Indic languages Ramesh et al. (2021). While the initial goal was to develop a translation model for Hindi to other Indic languagesGala et al. (2023), resource constraints prompted us to shift in our focus towards enhancing translation between English and Hindi and vice versa. In this report, we have also presented a preliminary analysis of translation between Hindi and Odia and vice versa. This involves improving the existing translation model's capabilities by incorporating additional intelligence and fine-tuning it with a limited dataset. The sentences used for analysis can be found at **the GitHub link.**

2 Related Work

IndicBART: A Pre-trained Model for Indic Natural Language Generation Dabre et al. (2022):

IndicBART is a multilingual sequence to sequence pre-trained model. It mainly utilizes the orthographic similarity between the Indic scripts to improve transfer learning between

similar Indic languages. Orthographic similarity refers to the resemblance or similarity in the written or script forms of languages. For example the languages Gujarati and Hindi seem similar when written. This similarity is exploited to convert 9 scripts into 1 base script. This increases the shared sub words in the vocabulary, and we observe that single script models enable better cross-lingual transfer while fine-tuning. The model IndicBART uses 6 encoder and decoder layers hidden and filter sizes of 1024 and 4096, respectively, and 16 attention heads. Due to the script independent nature of the model, IndicBART can be used for translating languages similar to Indian languages such as Sinhala and Nepali on which IndicBART has not been explicitly pre-trained.

Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic LanguagesDoddapaneni et al. (2023)

This paper aimed to improve the NLU (Natural Language Understanding) capabilities of Indic languages by making contributions along 3 important axes (i) monolingual corpora (ii) NLU testsets (iii) multilingual LLMs focusing on Indic languages. They created the largest monolingual corpora, IndicCorp, with 20.9B tokens covering 24 languages from 4 language families. They created a human- supervised benchmark, IndicXTREME, consisting of nine diverse NLU tasks covering 20 languages. Across languages and tasks, IndicXTREME contains a total of 105 evaluation sets, of which 52 are new contributions to the literature. They also trained IndicBERT v2, a state-of-the-art model supporting all the languages. Their first contribution towards serving low-resource languages was to release IndicCorp v2, the largest collection of corpora for languages spanning 4 Indic language families with 20.9 Billion tokens and 1.1 Billion sentences. IndicCorp not only supports more Indic languages but also improves upon the data for languages supported in existing collections. Their second contribution was IndicBERT v2, a multilingual LM pre trained on IndicCorp v2 and supporting the largest number of Indic languages compared to existing models such as XLM-R, MuRIL, and IndicBERT v1. The most important contribution was IndicXTREME, a human supervised benchmark containing evaluation sets for nine diverse tasks with each task covering 7-18 Indic languages per task. These include five classification tasks, two structure prediction tasks, one QA task, and one text retrieval task. Of the total 105 evaluation sets, summed across languages and tasks, 52 were newly created as a part of this benchmark. Through this work, they contributed towards all the fundamental requirements of developing Indic language technologies. These include IndicCorp v2, the largest pretraining corpus for 24 Indic languages, IndicBERT v2 a language model pre-trained on IndicCorp v2, and a holistic cross-lingual NLU benchmark, IndicXTREME, for 20 Indic languages. They even provided evidence for their design decisions and showed that pretraining models only on Indic languages resulted in much better performance on IndicXTREME.

IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages (Gala et al. (2023)):

This paper provides the data, evaluation metric and the model for Indic Machine Translation.This paper discusses about the parallel training data for all the 22 Indic languages mentioned in the Constitution of India, robust benchmarks for evaluating the models with respect to Indian contents and n-way translation for the Scheduled Languages.

The model is trained on BPCC (Bharat Parallel Corpus Collection) which contains 644K manually translated sentence pairs along with 126M newly added bi-text pairs, totalling 230M data instances. IndicTrans2 supports languages from four Indian language families-Indo-Aryan branch of the Indo-European family, Dravidian, Tibeto-Burman, and Austro-Asiatic. There are up to 12 major scripts spanning abugida, alphabetic, and abjad script types. For Pre-processing the data, all the numerals are converted to English numbers and extra spaces or special characters are removed. The punctuations are normalised and duplicate pairs of sentences are removed. The tag <dnt> was used for emails, URLs, dates etc. so that these terms do not change while translating. By using the IndicNLP library, rule-based script conversion was applied so that similar scripts could be represented as a single script. This resulted in a total of 5 scripts for all 22 languages which also contributed to simplifying the dataset for the model. To address the challenge of out-of-vocabulary (OOV) words in neural machine translation (NMT) systems, this work explores the effectiveness of subword-level tokenization. This approach segments text into smaller units, enhancing

robustness against OOV issues. The study leverages the byte-pair encoding (BPE) algorithm (Sennrich et al., 2016b) for training two separate tokenizers, facilitated by the SentencePiece library.

3 Methodology

In our project, we identified different kinds of biases present in the machine translation model and we will evaluate its performance using the BLEU score metric and human evaluation. we will fine-tune this model and evaluate its performance after fine-tuning and if required we will make changes to the model to get better results.

1. **Bias Assessment :** Initially, we identified the biases present in the machine translation model. These biases include relationship, gender and other forms of biases that could potentially affect the accuracy and fairness of translations produced by the model.
2. **Performance Evaluation Using BLEU Score:** For performance evaluation purposes we will use the BLEU (Bilingual Evaluation Understudy) score metric which is used to identify the similarity between the model translated output and the referenced translation. if the BLEU score is high it means our model translation is more similar to human translation.
3. **Human Evaluation:** Apart from quantitative evaluation, we are also using human evaluation which will provide qualitative feedback on the translations generated by the model. By human evaluations, we can assess the accuracy, fluency and overall quality of the translations that are done by the model. it offers valuable insights beyond numerical metrics.
4. **Fine-tuning and Model Adjustment:** In this step to reduce the biases in the model we will fine-tune the model parameter and train it on the task-specific data. if required we may make changes to the model architecture or training process to get better results.
5. **Post Fine-tuning Evaluation:** After fine-tuning the model we will re-evaluate the performance of the model using the same methodologies employed initially. this post-fine-tuning evaluation will help us to understand how much improvement has been reflected in the model, whether is it reducing the biases or not.

4 ANALYSIS AND RESULTS

4.1 Biases

1. **Gender-Based Tone Disparity:** In general, we observed that the tone of translation to Hindi was respectful even in the places where the tone lacked a little respect in English. However, we observed a discrepancy in the model's translation tone when the sentence is related to husband and wife. Precisely, when a sentence mentioned husband, the model translated it in a respectful tone in Hindi. However, when the same sentence mentioned wife, the model translated it in Hindi with a tone that lacked the same level of respect. This indicates a bias and the importance of ensuring fairness and equality in the model outputs.

   - **eng_Latn:** Return my money! Else you are doomed!!!
     hin_Deva:मेरे पैसे वापस कर दो! नहीं तो आप बर्बाद हो जाएँगे!!!
   - **eng_Latn:** My wife won't let me speak to my mother
     hin_Deva:मेरी पत्नी मुझे अपनी माँ से बात नहीं करने देगी।
   - **eng_Latn:** My husband won't let me speak to my mother
     hin_Deva:मेरे पति मुझे अपनी माँ से बात नहीं करने देंगे।

2. **Gender Bias:** IndicTrans2 mostly uses the masculine tone while translating irrespective of the location mentioned in sentence. The following structure was used where location=[park, mall, school, garden, library, parlour, salon, hospital]. Our expectation was in at least a few cases, irrespective of bias, we would observe the feminine version of the sentences. But it was not the case. We also observed two different inheritances i.e. old books and land or money and found the translations as shown below, which shows a sheer bias that prevails in our society.
   - **eng_Latn:** When I was young, I used to go to the park every day.
     hin_Deva:जब मैं छोटा था, मैं हर दिन पार्क जाता था।
   - **eng_Latn:** When I was young, I used to go to the mall every day.
     hin_Deva:जब मैं छोटा था, मैं हर दिन मॉल जाता था।
   - **eng_Latn:** When I was young, I used to go to the school every day.
     hin_Deva: जब मैं छोटा था, मैं हर दिन स्कूल जाता था।
   - **eng_Latn:** When I was young, I used to go to the garden every day.
     hin_Deva:जब मैं छोटा था, मैं हर दिन बगीचे में जाता था।
   - **eng_Latn:** When I was young, I used to go to the library every day.
     hin_Deva:जब मैं छोटा था, मैं हर दिन पुस्तकालयजाता था।
   - **eng_Latn:** When I was young, I used to go to the parlour every day.
     hin_Deva: जब मैं छोटा था, तो मैं हर दिन पार्लर जाता था।
   - **eng_Latn:** When I was young, I used to go to the salon every day.
     hin_Deva:जब मैं छोटा था, मैं हर दिन सैलूनजाता था।
   - **eng_Latn:** When I was young, I used to go to the hospital every day.
     hin_Deva: जब मैं छोटा था, तो मैं हर दिन अस्पताल जाता था।
   - **hin_Deva:**उसके पास बहुत सारी पुरानी किताबें हैं, जिन्हें उसने अपने दादा–परदादा से विरासत में पाया।
     eng_Latn: She has a lot of old books, which she inherited from her grandparents.
   - **hin_Deva:**उसके पास बहुत सारी सम्पति हैं, जिसे उसने अपने दादा–परदादा से विरासत में पाया।
     eng_Latn: He has a lot of wealth, which he inherited from his grandparents.
   - **hin_Deva:**उसके पास बहुत सारी ज़मीन है, जिसे उसने अपने दादा–परदादा से विरासत में पाया।
     eng_Latn: He owns a lot of land, which he inherited from his grandparents.
   - **ory_Orya**: ସେ ତାଙ୍କ ଦିଦିଙ୍କ ସହ ନୂଆ ଶାଢ଼ୀ କିଣିବା ପାଇଁ ବଜାରକୁ ଯାଇଥିଲେ। **hin_Deva**: वह अपनी बहन के साथ नई साड़ी खरीदने के लिए बाजार गई थी। [could have been masculine like other cases.]

4.2 Relationships

We noticed that in Hindi and other local Indic Languages, we use distinct terms for different relationships that falls under the categories of bigger umbrella terms such as sister-in-law, brother-in-law, uncle and aunt. Regional Indic Languages use specific words to precisely describe relationships. We believe that we can not create new words in English for such relationships but can add some intelligence to the model to capture some of the meaning from the context and translate accordingly. Here are the examples of translations that we found really interesting,

- **eng_Latn:** My brother's mother's daughter's only brother's wife's son's mother is my sister-in-law
  hin_Deva: मेरे भाई की माँ की बेटी की इकलौती भाई की पत्नी के बेटे की माँ मेरी साली है
- **hin_Deva:** मेरे भाई की माँ की बेटी के एकलौता भाई की पत्नी के बेटे की माँ मेरी भाभी है
  eng_Latn: The mother of my brother's mother's daughter's only brother's wife's son is my sister-in-law

- **eng_Latn:** My mother is my grandmother's daughter
  hin_Deva:मेरी माँ मेरी दादी की बेटी है।
- **eng_Latn:** My maternal grandmother cooks delicious food for us
  hin_Deva:मेरी नानी हमारे लिए स्वादिष्ट खाना बनाती हैं।
- **hin_Deva:** नानी माँ के लूसके हमेशा काम आते हैं
  ory_Orya: ଜେଜେମା। 'ଙ୍କ ପୋଷାକ ସବୁବେଳେ କାମରେ ଆସିଥାଏ।
  Maternal relationships were not given preference despite the context. Also while converting Indic-to-Indic maternal grandmother was translated as paternal grandmother.

4.3 Numbers

No Translation for numbers: Numbers are an essential part of the language. However, IndicTrans2 does not translate numbers or digits while translating. This leaves room for improvement. The model translates words such as "hundred" or "one" to the respective language but does not handle numerical values.

- **eng_Latn:** She gifted me a INR 100 thousand purse
  hin_Deva: उन्होंने मुझे 100 हजार रुपये का पर्स उपहार में दिया।
- **hin_Deva:** भोपाल गैस त्रासदी में लगभग ३००० लोगों की मौत हो गई
  ory_Orya: ଭୋପାଳ ଗ୍ୟାସ୍ ଦୁର୍ଘଟଣାରେ ପ୍ରାଯ 3000 ଲୋକ ପ୍ରାଣ ହରାଇଛନ୍ତି।
  Both Hindi and Odia numerals are translated into English digits. The words for numbers (hundred, thousand, 1.5, 2.5) were correctly translated. Since there is no concept of पौने, सवा in Odia, the terms were ignored while converting to Odia.

4.4 Homonyms

The model was able to identify when one word is used as a 'noun' or an 'adjective'. For example,

- **eng_Latn:** My sole is aching real bad
  hin_Deva:मेरी सोल बहुत बुरी तरह दर्द कर रही है।
- **eng_Latn:** The sole reason I left my job is to go for higher studies
  hin_Deva:मैंने अपनी नौकरी छोड़ने का एकमात्र कारण उच्च शिक्षा के लिए जाना है।

Since few words are similar in Odia and Hindi, the homonyms having same spelling were translated correctly in Indic to Indic but there were few exceptions:

- **ory_Orya**: ଏହା କାଳର ଶକ୍ତି ଏବଂ କାଳର ଅପରିହାର୍ଯ୍ଯତାର ଏକ ଅନୁସ୍ମାରକ। **hin_Deva**: यह समय की शक्ति और समय की अनिवार्यता की याद दिलाता है।

4.5 Proverbs and Idioms

We analysed many aspects of language translation such as, how the model translates proverbs and idioms in different languages. Ideally, it should recognize the presence of proverbs and not literally translate them to other languages. Instead, it should provide the actual meaning of the proverb/idiom in the target language or give the corresponding idiom in that language.

- **eng_Latn:** It's raining cats and dogs
  hin_Deva:बिल्लियों और कुत्तों की बारिश हो रही है
- **eng_Latn:** His wardrobe was at sixes and sevens
  hin_Deva:उनकी अलमारी छक्कों और सातों पर थी।

4.6 Miscellaneous

Unlike Hindi, Odia has no gender for verbs and things. Few words in Odia are same as Hindi but the pronunciation is different.

1. **Inconsistent translation of special characters:** IndicTrans2 removes duplicated special characters while pre-processing. The emojis not involving any alphabets [(^)3] are unaffected showing the expected behaviour. But the emojis involving alphabets[(O_O)] are translated individually by alphabets [(ଓ _ ଓ)] which is not the expected behaviour. The emoji also affected the quality of translation as can be seen from below:
   - **eng_Latn:** GOOD MORNING!!
     hin_Deva: अच्छी सुबह!!
   - **eng_Latn:** Do Visit The Grand Sale At Orion Mall
     hin_Deva: ओरियन मॉल की ग्रैंड सेल में जरूर जाएँ।
   - **eng_Latn:** Do Visit The Grand Sale At Orion Mall :)
     hin_Deva: ओरियन मॉल में भव्य बिक्री पर अवश्य जाएँ:)

   While translating Indic to Indic, the model uses single and double quotes interchanagably but in Indic language its not the case.

   - **hin_Deva**: उसकी क्या बात कर रहे हो, वह तो मेरे खून का प्यासा हो गया है। **ory_Orya**: ଆପଣ କ "ଣ କହୁଛନ୍ତି, ସେ ମୋ ରକ୍ତ ପାଇଁ ତୃଷାର୍ତ୍ତ।
2. **Case of the sentence:** The model behaves inconsistently when the capitalization is altered. For some input cases, the translation was the same irrespective of the case of the sentence but a few most commonly used sentences as shown below were translated differently when the capitalization was toggled.
   - **eng_Latn:** wE WISH YOU A mERRY cHRISTMAS AND A hAPPY nEW yEAR
     hin_Deva: हम आपको एक मेरी क्रिसमस और एक खुशहाल नए साल की कामना करते हैं
     eng_Latn: We wish you a Merry Christmas and a Happy New Year
     hin_Deva: हम आपको क्रिसमस और नए साल की शुभकामनाएँ देते हैं।
3. **Naming words:**
   - Food items that have an English and Hindi name like Indian mashed potatoes, Naan Roti, and fritters could not be correctly translated to "Aloo Bharta", "Naan Roti", "Pakoda". But dishes like rasagulla, and chole were correctly translated as per the syllabi. There are multiple food items which have similar names in both Hindi and Odia. Even such names were not translated correctly and were mostly translated wrongly based on similar phonetics.
     - **ory_Orya:** ଆମେ ଆଳୁ ଭର୍ତ୍ତା ଖାଇଲୁ।
       hin_Deva: हमने आलू का आटा खाया।
     - **hin_Deva:** पानी पुरी खाना चाहते हैं?
       eng_Latn: Want to drink water?
   - Locations like "Goregaon East" were literally translated to Hindi instead of keeping the name as it is.
4. **Indian English:** There are many peculiar phrases in Indian English and Hindi (and in every other language) that are used generally in almost all contexts but do not have a particular meaning. The model translates the words literally and thus loses the essence in which it was conveyed. It is unable to translate "out of station", "please revert back", etc.

- **eng_Latn**: He is out of station.
  hin_Deva: वह स्टेशन से बाहर है।
- **hin_Deva:** नानी माँ के लूसके हमशा काम आते हैं
  eng_Latn: Grandmother's clothes always come in handy.

5. **Felicitations** : The model is still not good at conveying wishes.
   - **eng_Latn:** Best of luck for your exams!
     hin_Deva: परीक्षा के लिए आपको बहुत–बहुत बधाई!
   - **eng_Latn:** We are cordially inviting you to grace us with your presence as we, Abhisekh and Sweta, unite in marriage. Let's create unforgettable memories together on 4th April, 2024.
     hin_Deva:
     हम आपको अपनी उपस्थिति के साथ हमें खुश करने के लिए आमंत्रित कर रहे हैं क्योंकि हम, अभिषेक और श्वेता, शादी में एकजुट होते हैं। आइए 4 अप्रैल, 2024 को एक साथ अविस्मरणीय यादें बनाएं।
6. Some words like पुस्तकालय , सभागार, अंक were incorrectly transliterated to **English** form i.e. instead of writing Odia word for library, auditorium and point, the model typed the English words in Odia letters. The same was observed while translating Odia to Hindi.
   - **hin_Deva:** जब मैं छोटा था, मैं हर दिन पुस्तकालय जाता था।
     ory_Orya: ଯେତେବେଳେ ମୁଁ ଛୋଟ ଥିଲି, ମୁଁ ପ୍ରତିଦିନ ଲାଇବ୍ରେରୀକୁ ଯାଉଥିଲି।
7. Given that in Odia ୟ, ଯ, ଜ are different, the model does not output the letter ୟ properly.
   - ବକେଯ , ପ୍ରାଯ, ଭାରତୀଯମ୍ାନେ...
8. **Unsynced chronology of sentence:** The model wrongly translated the following denoting that it cannot grasp the temporal context.
   - **hin_Deva:** अगर आप उस समय मुझसे मिलते तो हम बाहर खाना खाने जाते।
     ory_Orya: ଯଦି ଆପଣ ମୋତେ ସେହି ସମଯରେ ଭେଟିଥାନ୍ତେ, ତେବେ ଆମେ ରାତ୍ରିଭୋଜନ ପାଇଁ ବାହାରକୁ ଯାଉଥିଲୁ। [Used both if clause and past clause instead of future]

Conclusion

We have analysed various test sentences for the IndicTrans2 model and it works really well in translating the majority of the sentences. This achievement fosters greater communication and accessibility within India's diverse linguistic landscape. English speakers can now effectively interact with speakers of other scheduled languages, and vice versa. Additionally, the model supports communication between speakers of various Indian languages themselves. However, there are a few exceptions as discussed in the Analysis and Results section. Thus there is room for improvement for the model. Translation of numbers, relationships, proverbs and idioms are a few aspects that can be worked on. Further, a similar analysis can be carried out for the generative model.

Contributions

The authors have equally contributed to the writing of the report. Other tasks performed for this report were researching and analyzing the models by various test cases. The team had discussed and brainstormed the topics under which the model could have performed unsatisfactorily. All the members have examined the English to Hindi and vice versa test cases collectively.

1. **Diksha Seth:**
   - Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic LanguagesDoddapaneni et al. (2023)

- IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian LanguagesKakwani et al. (2020)
- Attempted analysis of Hindi to Punjabi and vice-versa.

2. **Sachin Kumawat:**
   - Working of Evaluation Metrics- BLEU, chrF
   - Exploring how to run models like IndicBERTv1, IndicTrans2, etc
3. **Sofia Sunam:**
   - IndicBART: A Pre-trained Model for Indic Natural Language Generation Dabre et al. (2022)
   - IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages Gala et al. (2023)
   - Analysis of Odia to Hindi and vice-versa.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL `https://aclanthology.org/2022.naacl-main.223`.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.145. URL `http://dx.doi.org/10.18653/v1/2022.findings-acl.145`.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages, 2023.

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

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh Khapra, and Pratyush Kumar. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. pp. 4948–4961, 01 2020. doi: 10.18653/v1/2020.findings-emnlp.445.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Deepak Kumar, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar:

The largest publicly available parallel corpora collection for 11 indic languages. *CoRR*, abs/2104.05596, 2021. URL https://arxiv.org/abs/2104.05596.