TRANSLATION FOR INDIAN LANGUAGES

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayak 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, Mitesh M. Khapra

SUMMARY

- ***Samanantar:*** **Largest publicly available parallel translation corpus for 11 Indian languages**
- ***IndicTrans:*** **Open-source models supporting translation between Indian languages and English**
- **We release Indic-Indic parallel sentence corpus**
- **We release human annotation scores of the *Samanantar* corpus**

What is Machine Translation?

The goal of AI4Bharat is to build language technologies for all Indian languages

एआई4भारत का लक्ष्य सभी भारतीय भाषाओं के लिए भाषा प्रौद्योगिकियों का निर्माण करना है

What is missing for Indian Languages?

Large Scale MT training data

IN-22 specific Innovations in Models

Our Approach

1. **Mine data at scale**
2. **Train MT models exploiting language similarities**

1. Mining Approach

Data Sources

Port of Spain, July 25: Axar Patel's blitzkrieg with the bat helped Team India pull off a sensational two-wicket win over West Indies in the second one-day international at Port of Spain on Sunday (July 24). Patel - who slammed his ODI career's maiden half-century - held on to his nerves in the death overs and guided the Men In Blue to a series-clinching victory in Trinidad and Tobago.

https://mykhel.com

पोर्ट ऑफ स्पेन, जुलाई 25। वेस्टइंडीज के खिलाफ दूसरे वनडे में कप्तान की भूमिका करने वाले अक्षर पटेल ने पूर्व भारतीय कप्तान महेंद्र सिंह धोनी का एक रिकॉर्ड तोड़ दिया है। दरअसल, अक्षर पटेल ने सातवें नंबर पर आकर सफलतम स्व पेज में भारत की ओर से एक पारी में सबसे ज्यादा छक्के लगाने का रिकॉर्ड अपने नाम कर लिया है। कल की पारी में अक्षर पटेल ने 3 छक्के लगाए थे। अक्षर से पहले एमएस धोनी ने 2005 में जिम्बाब्वे के खिलाफ अपनी पारी में 3 छक्के लगाए थे।

https://hindi.mykhel.com

Nearest Neighbour search

37.4M Sentence Pairs

2. Language Similarity

Marathi → भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला

Marathi Segmented → *भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला*

Hindi → भारत के स्वतंत्रता दिवस के अवसर पर अमरिका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया

We use script unification to leverage language similarity

3. Training Large NMT Models

Curate large scale NMT training data

Script Unification

Train multilingual NMT models

हुं અહીં છું → हुं अहीं छुं → IndicTrans → I am here

4. Resources

49.7M English Centric Corpus

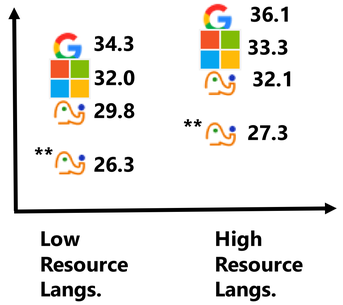
83.4M Indic-Indic Corpus

434M Parameter Model

Support for 11 Indian languages

Results

Training with Samanantar improves BLEU score



OUR PLAN AHEAD

- **Support 22 Indian languages**
- **Improve translation quality**
- **Creating efficient models for deployment**
- **Create benchmarks for IN-22**

ACKNOWLEDGEMENTS

We would like to thank the Nilekani Philanthropies for their generous grant which helped in setting up the "Nilekani Centre at AI4Bharat, IIT Madras" to support our students and research staff, as well as data and computational requirements. We would like to thank The Ministry of Electronics and Information Technology for its grant to support the creation of datasets and models for Indian languages under its ambitious Digital India Bhashini project. We would also like to thank the Centre for Development of Advanced Computing, India (C-DAC) for providing access to the Param Siddhi supercomputer for training our models. Lastly, we would like to thank Microsoft for its grant to create datasets and tools for Indian languages.

The focus of AI4Bharat, an initiative of IIT Madras, is on building open-source language AI for Indian languages, including datasets, models, and applications.

https://ai4bharat.iitm.ac.in/indic-trans
https://github.com/AI4Bharat/indicTrans
Contact: Gowtham Ramesh, Sumanth Doddapaneni