ENG2TEL Transalatron:English to Telugu Translation Project -Bridging Linguistic Gaps

Alokam Gnaneswara Sai & Onteru Prabhas Reddy & Pallekonda Naveen Kumar
Indian Institute of Science
Bengaluru, KA, India
{gnaneswaras,prabhasreddy,kpnaveen}@iisc.ac.in

1 Introduction

In recent years, language translation has made significant strides, thanks to advancements in deep learning. These improvements have greatly enhanced translation accuracy and fluency, especially for widely spoken languages like English and French. However, low-resource Indian languages like Telugu, spoken by an estimated 96 million people in India, have not seen the same progress comparatively.

The main challenge in developing effective translation systems for languages like Telugu is the lack of enough data and resources. Also, there aren't many researchers focusing on these languages. This lack of attention has led to a big difference in translation quality as compared to other popular languages.Even current state-of-the-art models like GPT-3 lacks understanding of the languages like Telugu. Telugu speakers often struggle to find accurate translation tools, making it hard for them to access information.

Given Telugu's importance in Indian culture, it's crucial to bridge this translation gap and provide better access to information for Telugu speakers. That's where our project comes in. We're dedicated to improving translation from English to Telugu. By using the "Samanantar"Gowtham Ramesh1 (2023) dataset, collected in 2023, we aim to develop more accurate and helpful translation models.

Our goal is to improve communication for Telugu speakers, making it easier for them to access information. By providing better translation tools, we hope to bring languages closer together, promote cultural exchange, and encourage meaningful interactions in our increasingly connected world.

2 Related Work

Recent advancements in Indian language translation, as evidenced by models like BLOOMBigScience Workshop: Teven Le Scao (2023), IndicBARTJay Gala (2022), and IndicTrans2Jay Gala (2023) , have shown promising results in handling multiple Indian languages simultaneously. However, their broad focus on accommodating diverse languages may inadvertently compromise the accuracy of translation, particularly for languages like Telugu. The inherent complexity and unique linguistic features of each language pose significant challenges that may not be fully addressed by a one-size-fits-all approach.

Moreover, the scarcity of high-quality datasets further exacerbates the challenges in developing accurate translation models for Telugu and other low-resource languages. Recognizing this gap, recent initiatives have emerged to address the need for specialized resources tailored specifically for Telugu language processing tasks.

For instance, the "Namapadam"Arnav Mhaske (2023) dataset has been instrumental in advancing named entity recognition (NER) tasks in Telugu and other low resource languages. By providing meticulously annotated Telugu text, this dataset enables researchers to train and evaluate NER models effectively, thereby enhancing the performance of language processing tasks that rely on entity recognition.

Additionally, the "Samanantar" dataset has emerged as a valuable resource for sentence-level translation between English and Telugu. With parallel English-Telugu sentences of more than 4.6M sentences, this dataset facilitates the development of more accurate and contextually relevant translation models. By providing clean and aligned data, "Samanantar" enables researchers to overcome the challenges posed by data scarcity, paving the way for improved language processing solutions tailored specifically for Telugu.

3 Results

In our examination of existing translation models, we assessed the performance of IndicTrans2 and IndicBART for English to Telugu translation. IndicTrans2 achieved a highest BLEU score of 19.4 for translating from English to Telugu and 42.3 for Telugu to English translation on IN22-Gen Evaluation set.But whereas other multilingual models like IT1,Google,Azure has less BLEU score comparatively.

Meanwhile, IndicBART exhibited a BLEU score of **29** for English to Telugu translation on the WAT 2021 test set. , accompanied by a ROUGE score of **14**. These findings provide insights into the current landscape of translation capabilities for Telugu and serve as a baseline for evaluating the efficiency of our proposed methods in enhancing translation quality.

Models	**IT1**	**N1.2**	**N54**	**IT2**	**Goog**	**Az**
BLEU Scores	15.5	15.1	17.1	19.4	17.7	17.7

Table 1: BLEU scores on the IN22-Gen Evaluation set for English to Telugu.

Models	**IT1**	**N1.2**	**N54**	**IT2**	**Goog**	**Az**
BLEU Scores	12.0	9.80	10.5	14.1	13.4	13.8

Table 2: BLEU scores on the IN22-Conv Evaluation set for English to Telugu.

Conclusion

In conclusion, our study has highlighted the issue of inadequate translation solutions for low-resource languages like Telugu, which poses significant challenges for effective communication and access to information for Telugu speakers.

To address this challenge, our approach involves experimenting with various encoder-decoder models and implementing fine-tuning techniques tailored specifically to the linguistic features of Telugu. By adapting existing methodologies to the unique characteristics of Telugu, our goal is to enhance translation accuracy and fluency for this language.

Moving forward, our evaluation will involve comparing the performance of our fine-tuned models with existing translation results for other popular languages. Through rigorous testing and analysis, we aim to assess the effectiveness of our approach in improving translation capabilities for Telugu, thereby contributing to the advancement of language processing solutions for under-resourced languages.

Contributions

- Prabhas Reddy helped in identifying and refining the research topic, and has read research papers to find relevant topics for the project direction.

- Gnaneswara Sai explored and verified different datasets used in the research, ensuring their quality for analysis and experimentation. Explored the Samanantar dataset and helped in reading research papers.
- Naveen Kumar has read relevant research papers in the domain of language translation, contributing to the literature review and understanding of existing approaches and techniques.

References

Sumanth Doddapaneni Mitesh M. Khapra Pratyush Kumar Rudra Murthy V Anoop Kunchukuttan Arnav Mhaske, Harshit Kedia. In *Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages*, volume 14. arXiv preprint arXiv:2212.10168, 2023. URL https://arxiv.org/pdf/2212.10168.pdf.

Christopher Akiki Ellie Pavlick Suzana Ilić Daniel Hesslow Roman Castagné Alexandra Sasha Luccioni François Yvon Matthias Gallé Jonathan Tow Alexander M. Rush Stella Biderman Albert Webson Pawan Sasanka Ammanamanchi Thomas Wang Benoît Sagot Niklas Muennighoff Albert Villanova del Moral Olatunji Ruwase Rachel Bawden Stas Bekman Angelina McMillan-Major Iz Beltagy Huu Nguyen Lucile Saulnier Samson Tan Pedro Ortiz Suarez Victor Sanh Hugo Laurençon Yacine Jernite Julien Launay Margaret Mitchell Colin Raffel Aaron Gokaslan Adi Simhi Aitor Soroa Alham Fikri Aji Amit Alfassy Anna Rogers Ariel Kreisberg Nitzav Canwen Xu Chenghao Mou Chris Emezue Christopher Klamm Colin Leong Daniel van Strien David Ifeoluwa Adelani Dragomir Radev Eduardo González Ponferrada Efrat Levkovizh Ethan Kim Eyal Bar Natan Francesco De Toni Gérard Dupont Germán Kruszewski Giada Pistilli Hady Elsahar Hamza Benyamina Hieu Tran Ian Yu Idris Abdulmumin Isaac Johnson Itziar Gonzalez-Dios Javier de la Rosa Jenny Chim Jesse Dodge Jian Zhu Jonathan Chang Jörg Frohberg Joseph Tobing Joydeep Bhattacharjee Khalid Almubarak Kimbo Chen Kyle Lo Leandro Von Werra Leon Weber Long Phan Loubna Ben allal Ludovic Tanguy Manan Dey Manuel Romero Muñoz Maraim Masoud María Grandury Mario Šaško Max Huang Maximin Coavoux Mayank Singh Mike Tian-Jian Jiang Minh Chien Vu Mohammad A. Jauhar Mustafa Ghaleb Nishant Subramani Nora Kassner Nurulaqilla Khamis Olivier Nguyen Omar Espejel Ona de Gibert Paulo Villegas et al. (293 additional authors not shown) BigScience Workshop: Teven Le Scao, Angela Fan. In *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*, volume 73. arXiv preprint arXiv:2211.05100, 2023. URL https://arxiv.org/pdf/2211.05100.pdf.

Aravinth Bheemaraj Mayank 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 Shantadevi† Khapra1 Gowtham Ramesh1, Sumanth Doddapaneni1. In *ISamanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages*, volume 16. arXiv preprint arXiv:2104.05596, 2023. URL https://arxiv.org/ftp/arxiv/papers/2104/2104.05596.pdf.

Raghavan AK Varun Gumma Sumanth Doddapaneni Aswanth Kumar Janki Nawale Anupama Sujatha Ratish Puduppully Vivek Raghavan Pratyush Kumar Mitesh M. Khapra Raj Dabre Anoop Kunchukuttan Jay Gala, Pranjal A. Chitale. In *IndicBART: A Pre-trained Model for Indic Natural Language Generation*, volume 15. arXiv preprint arXiv:2109.02903, 2022. URL https://arxiv.org/pdf/2109.02903.pdf.

Raghavan AK Varun Gumma Sumanth Doddapaneni Aswanth Kumar Janki Nawale Anupama Sujatha Ratish Puduppully Vivek Raghavan Pratyush Kumar Mitesh M. Khapra Raj Dabre Anoop Kunchukuttan Jay Gala, Pranjal A. Chitale. In *IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages*, volume 95. arXiv preprint arXiv:2305.16307, 2023. URL https://arxiv.org/ftp/arxiv/papers/2305/2305.16307.pdf.