Multilingual Machine Translation for Code-Mixed Hinglish and Tanglish Sentences to English: Leveraging NMT for Informal Communication

1. Introduction:

In multilingual societies like India, informal conversations often contain code-mixing, especially in written digital communications such as text messages, social media posts, and emails. Code-mixing occurs when users combine two or more languages, particularly when a user is fluent in multiple languages but chooses to combine elements of each during communication. In India, this phenomenon is seen with several languages, but we'll be focusing on Hindi and Tamil, which are frequently mixed with English. This leads to hybrid languages like Hinglish (Hindi-English) and Tanglish (Tamil-English), where Hindi or Tamil words are written in Roman script, combined with English words.

Despite this linguistic norm, machine translation systems often fail to handle code-mixed languages effectively, especially in informal chat-based settings where grammatical consistency is rare. The lack of training data and resources further exacerbates the issue.

This project aims to try, build, test and fine tune existing various neural machine translation (NMT) models capable of accurately translating code-mixed text in Hinglish and Tanglish into both pure Hindi, Tamil, and subsequently into English. We will compare various models such as mBART, mT5, IndicTrans2 for this task using existing corpora and introduce synthetic data generation methods to fill in the resource gaps if required.

2. Motivation

Given the rise of social media, informal chat systems, and texting as primary communication modes, there is an increasing need to develop machine translation systems that can handle real-world multilingual, code-mixed texts. A machine translation system capable of processing Hinglish and Tanglish effectively can enhance the overall user experience for multilingual communication platforms such as WhatsApp, Facebook Messenger, Twitter, and Instagram. This system can find use in:

a. Chat translation tools for multilingual customer support.
b. Cross-language sentiment analysis and news aggregation in financial and social media.
c. Language learning applications that can switch between informal and formal contexts.
d. Conversational AI for generating and interpreting responses in code-mixed formats.

3. Existing Work & Literature Review:

Recent research has explored multilingual neural machine translation for low-resource languages, including methods for code-mixed data processing. Several notable contributions that align with our project are:

1. IndicTrans2 (AI4Bharat): A multilingual translation model designed specifically for Indian languages, focusing on improving performance for languages like Tamil and Hindi by leveraging shared linguistic structures.

2. Hinglish Translation Systems: Research from CALCS2021 and WMT2022 has demonstrated success in generating synthetic code-mixed data to train models for English-Hinglish translation. Methods such as backtranslation, bilingual embeddings, and curriculum learning have been used to improve performance in this paper.

3. Multimodal NMT for Dravidian Languages: Work by Chakravarthi et al. introduces phonetic transcription and multimodal features to improve translation for Dravidian languages. This shows significant promise for Tanglish, which lacks adequate translation resources (data).

4. Handling Code-Mixing in Low-Resource Languages: Researchers like Jawahar et al. (2021) have introduced methods such as curriculum learning and data augmentation to overcome the lack of labeled datasets.

These works provide a foundation for the project by showing how NMT models can be adapted for informal code-mixed translations using synthetic data, back transliteration, and pretrained and modified multilingual transformers like mBART and mT5.

4. Project Outline:

This project will address the following research questions:
How can we accurately translate Hinglish and Tanglish texts into formal Hindi, Tamil, and English using existing NMT and transliteration models?
How do code-mixed language translation challenges vary between high-resource (Hindi) and low-resource (Tamil) languages?
What improvements can be made through the introduction of synthetic code-mixed datasets and fine-tuning of multilingual models?

5. Methodology:

1. Data Collection & Preprocessing:

We'll use publicly available datasets for Hinglish and Tamil-English parallel corpora from Samanantar, PHINC, IndicMT-Eval and make our own datasets by scraping code-mixed comments from YouTube, Instagram and other social media websites.

2. Building Model Pipeline:

The first half of the pipeline involves translating the code-mixed texts using transliteration models to get the proper native text. The second half involves converting the native text to proper English using popular neural language translation models. We will experiment with mBART, mT5, IndicTrans2, and other models in the second half for native text translations.

3. Model Training:

We train the selected models on Hinglish-to-Hindi/English and Tanglish-to-Tamil/English translations using parallel corpora and datasets obtained through web scraping. We fine-tune these models on synthetic code-mixed data using curriculum learning approaches.

4. Evaluation:

Performance will be evaluated using metrics such as BLEU, ROUGE, and METEOR scores. We'll compare results across models to assess the impact of code-mixing on translation quality for these two languages. We will also analyze which combination of transliteration and translation models work best for Hindi and Tamil.

6. Expected Challenges:

Pre-processed Data Scarcity for Tamil: While Hindi-English has abundant resources, Tamil-English (Tanglish) translations face some data scarcity. We could address this using synthetic data generation methods.

Informal Language Variability: The informal nature of code-mixed chats may pose challenges due to inconsistent spelling and sentence structures that don't follow strict grammar.

7. Conclusion:

This project will survey and test the development of machine translation systems for code-mixed languages, particularly for informal communication settings. By focusing on Hinglish and Tanglish, we aim to address a critical gap in multilingual NMT for low-resource languages, providing valuable insights into both synthetic data generation and model fine-tuning for other real-world applications.

8. References:

(1) Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

(2) Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2021. IITP-MT at CALCS2021: English to Hinglish Neural Machine Translation using Unsupervised Synthetic Code-Mixed Parallel Corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 31–35, Online. Association for Computational Linguistics.

(3) Vibhav Agarwal, Pooja Rao, and Dinesh Babu Jayagopi. 2021. Hinglish to English Machine Translation using Multilingual Transformers. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 16–21, Online. INCOMA Ltd.

(4) Jawahar, Ganesh & Nagoudi, El Moatez Billah & Abdul-Mageed, Muhammad & Lakshmanan, Laks. (2021). Exploring Text-to-Text Transformers for English to Hinglish Machine Translation with Synthetic Code-Mixing. 10.48550/arXiv.2105.08807.

(5) Kumar, Somnath & Balloli, Vaibhav & Ranjit, Mercy & Ahuja, Kabir & Ganu, Tanuja & Sitaram, Sunayana & Bali, Kalika & Nambi, Akshay. (2024). Bridging the Gap: Dynamic Learning Strategies for Improving Multilingual Performance in LLMs. 10.48550/arXiv.2405.18359.

(6) Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus Creation for Sentiment Analysis in Code-Mixed Tulu Text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

(7) I. Jadhav, A. Kanade, V. Waghmare, S. S. Chandok and A. Jarali, "Code-Mixed Hinglish to English Language Translation Framework," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2022, pp. 684-688, doi: 10.1109/ICSCDS53736.2022.9760834.

(8) Chakravarthi, Bharathi. (2023). Detection of homophobia and transphobia in YouTube comments. International Journal of Data Science and Analytics. 18. 1-20. 10.1007/s41060-023-00400-0.

(9) Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Jayapal, Sridevy S, Mihael Arcan, Manel Zarrouk, and John P McCrae. 2019. Multilingual Multimodal Machine Translation for Dravidian Languages utilizing Phonetic Transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63, Dublin, Ireland. European Association for Machine Translation.

(10) Akshat Gahoi, Jayant Duneja, Anshul Padhi, Shivam Mangale, Saransh Rajput, Tanvi Kamble, Dipti Sharma, and Vasudev Varma. 2022. Gui at MixMT 2022 : English-Hinglish : An MT Approach for Translation of Code Mixed Data. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1126–1130, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.