CoMeT: Towards Code-Mixed Translation Using Parallel Monolingual Sentences

A project report submitted as part of the requirements for the course titled Advanced NLP - Monsoon 2022

By
Team Number : 22
Team Name : The Duo
Gopichand Kanumolu (2021701039), gopichand.kanumolu@research.iiit.ac.in
Lokesh Madasu (2021701042), lokesh.madasu@research.iiit.ac.in

Mentor: Tanvi Kamble

International Institute of Information Technology
Hyderabad

16, November, 2022

Contents

1 Introduction

1.1 Task Description

Code-mixing is using words from two or more languages in the same conversation. The code-mixed text contains words from different languages. Due to the increase in social media usage, people are communicating informally, like expressing their emotions/feelings in the form of emojis, slang, and code-mixed languages. We often find code-mixed text on social media platforms such as Twitter, Facebook, Instagram etc.

1.2 Problem Statement

Code-Mixed data is massively used by multilingual people in their everyday lives. The major problem with code-mixed text is, if a person has to understand the meaning of text, he/she must know the different languages used in the text. Even knowing one language will not help in this case. To address this problem, we need a machine translation system that can able to understand the code-mixed text and translate it to the given target language. Building a code-mixed machine translation system is not a trivial task. So many challenges are involved like the system should be able to capture/understand the language-independent properties from the text and the code-mixed languages do not follow a prescriptively defined structure. In this project we attempt to build a machine translation system that translates the English sentence into codemixed language Hinglish (combination of words from Hindi and English). To achive this we will replicate/reproduce the results in the baseline paper ”CoMeT: Towards Code-Mixed Translation Using Parallel Monolingual Sentences.” and we will also explore different methodologies and datasets and will do detailed error analysis.

1.2.1 Why it is important?

Recent reports say that majority of the online textual data is code-mixed, given the rapid spread of code-mixed usage across the globe. So far, the state-of-the-art NLP models and applications are developed only in monolingual cases. All the NLP applications such as Chatbot, Speech recognition, Machine translation, Natural language generation, Text summarization, and Spelling correction were developed using monolingual data. Due to the rapid usage of code-mixed languages around the world, it is very important to create NLP models for code-mixed data.

2 Dataset

2.1 Hindi-English CodeMix Dataset

We use the CALCS Shared Task English-Hinglish dataset that is being used in the research paper that we are implementing. The dataset can accessed using the URL, https://ritual.uh.edu/lince/datasets. In the following sections we present the dataset statistics and plots.

2.1.1 Dataset Statistics and Exploratory Data Analysis

2.1.2 Dataset Statistics

Field	English	Hinglish
Number of sentences in train data	8060	8060
Number of sentences in validation data	942	942
Number of sentences in test data	960	960
Minimum Number of Tokens	1	1
Maximum Number of Tokens	287	370
Average Number of Tokens	12.37	12.6

Table 1: dataset statistics

Field	Value
Duplicate English-Hindi Sentence Pairs in train set	519
Duplicate English-Hindi Sentence Pairs in validation set	12
Duplicate English Sentences in test_set	182
Number of Hindi tokens in target sentences	76364
Number if English tokens in target sentences	24269
Number of 'Other' tokens in target sentences	13730

Table 2: other statistics

2.1.3 Plot of Number of Tokens in English Sentence Vs Number of Sentences

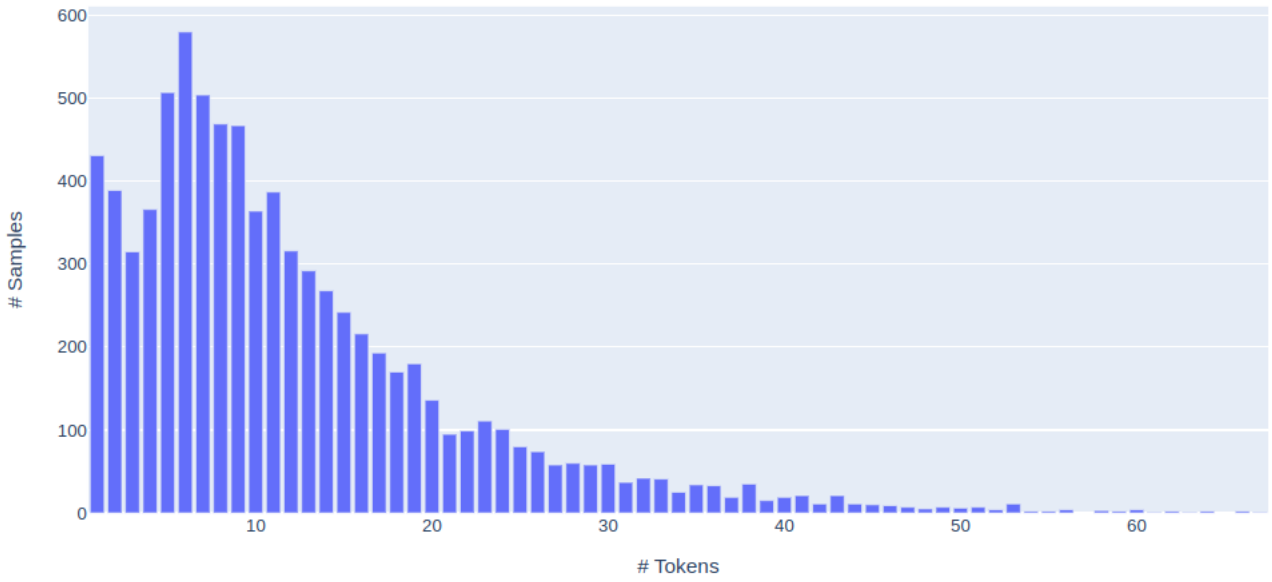

Figure 1: Number of Tokens in Text Vs Frequency (English)

2.1.4 Plot of Number of Tokens in Hinglish Sentence Vs Number of Sentences

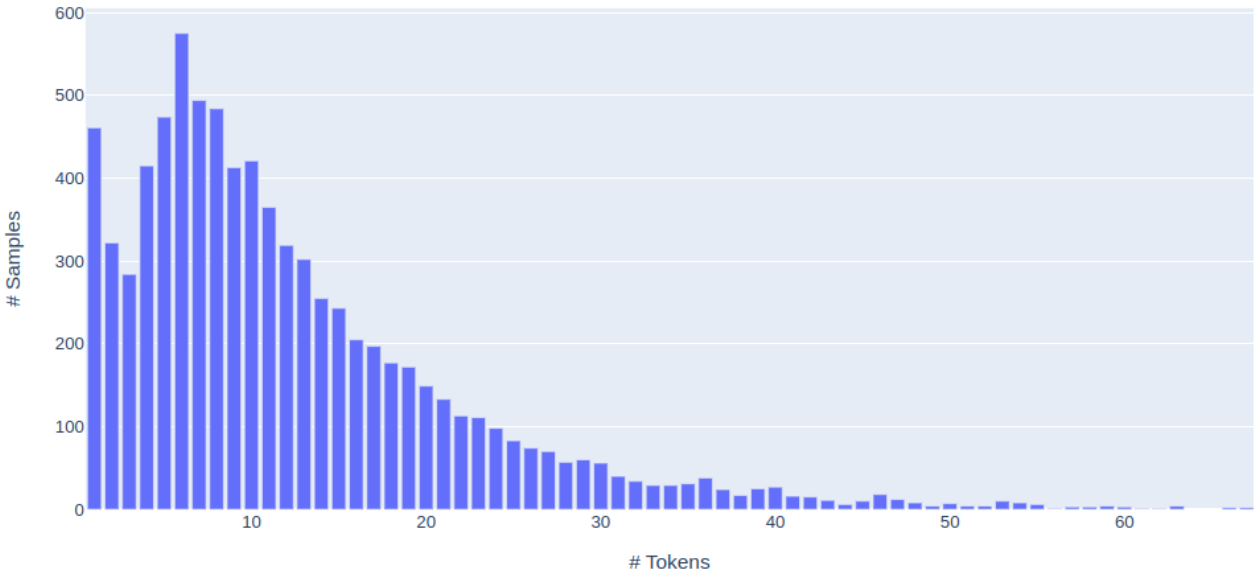

Figure 2: Number of Tokens in Text Vs Frequency (Hinglish)

2.2 Telugu-English CodeMix Dataset

We use the Telugu-English codemix data that is used for sentiment analysis task by the authors of the research paper "Sentiment Analysis in Code-Mixed Telugu-English Text with Unsupervised Data Normalization". The dataset contains 19,868 Telugu code mixed sentences. Each word in a sentence is given a label with respect to its language. For example, If a word belongs to Telugu, it its corresponding language tag is 'te'. If it belongs to Hindi it is marked 'hi' and if it is named entity it is marked 'ne' and words that do not fall under these categories is given a universal label 'univ', for example punctuations and special symbols etc that are language independent. We choose the sentences having at least one English and one Telugu word. After applying the filter we end up getting 16,249 sentences. Since, we have the language tags for each word in all the sentences, to determine the Code Mixing Index (CMI) we use the following formula to determine the amount of code mix present in the dataset. where, w_i is the number

$$CMI = \begin{cases} 100 \times [1 - \frac{max\{w_i\}}{n-u}] & n > u \\ 0 & n = u \end{cases}$$

Figure 3: Code Mixing Index(CMI)

of words tagged with a particular language tag, $maxw_i$ represents the number of words of the most prominent language, n is the total number of tokens, u represents the number of language independent tokens, in our case these would be the words marked as "univ". We report that the dataset has average CMI of 28.31. Since these sentences are Code Mix Telugu-English sentences, that are used for sentiment analysis task, we don't have its equivalent English sentence. To get their equivalent English sentences, we first transliterated the sentences that are in roman script to devanagari script and then we use google translate

to translate those sentence to English. We then form a training set with source as English sentence and target as code mixed sentence which is in devanagari.

3 Model Implementation

3.1 Existing Approaches (used in baseline paper)

mBART is a sequence-to-sequence denoising auto-encoder pretrained on large-scale monolingual corpora in many languages using the BART objective. mBART is one of the first methods for pretraining a complete sequence-to-sequence model by denoising full texts in multiple languages, while previous approaches have focused only on the encoder, decoder, or reconstructing parts of the text.

For machine translation task, the proposed approach in the research paper use mBART for translating English sentences to code-mixed sentences. Two strategies were used to achieve good performance. One is, the mBART-en model takes English sentences as input and generates Hinglish sentences as output. Second, the mBART-hien model takes both English sentences and their corresponding Hindi translations as input and generates Hinglish sentences as output.

To train our systems efficiently, we prune mBART's vocabulary by removing the tokens which are not present in the provided dataset or the dataset r leased by Kunchukuttan et al. (2018) which contains 1,612,709 parallel sentences for English and Hindi.

3.1.1 mBART-en

We fine-tune mBART on the train set, feeding the English sentences to the encoder and decoding Hinglish sentences. We use beam search with a beam size of 5 for decoding.

3.1.2 mBART-hien

We fine-tune mBART on the train set, feeding the English sentences along with their parallel Hindi translations to the encoder and decoding Hinglish sentences. For feeding the data to the encoder, we concatenate the Hindi translations, followed by a separator token ##, followed by the English sentence.

3.1.3 Model Parameters

Field	Value
Loss	label_smoothed_cross_entropy
Optimizer	adam
Learning Rate	3.00E-05
Dropout	0.3
Attention Dropout	0.1
Epochs	40
Max token length	312

Table 3: Model Paremeters

3.2 Additional Approaches (not used in baseline paper)

3.2.1 mT5

Multilingual T5 (mT5) is a massively multilingual pretrained text-to-text transformer model, trained following a similar recipe as T5. mT5 is pretrained on the mC4 corpus, covering 101 languages. We use mT5 model for following language pairs, English-Hindi and English-Telugu and use it for machine translation task.

3.2.2 IndicBART

IndicBART is a multilingual, sequence-to-sequence pre-trained model focusing on Indic languages and English. It currently supports 11 Indian languages and is based on the mBART architecture. You can use IndicBART model to build natural language generation applications for Indian languages by finetuning the model with supervised training data for tasks like machine translation, summarization, question generation, etc. The model is much smaller than the mBART and mT5(-base) models, so less computationally expensive for finetuning and decoding. Trained on large Indic language corpora (452 million sentences and 9 billion tokens) which also includes Indian English content.

4 Evaluation Metrics

4.1 BLEU

BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate's translation of the text to one or more reference translations. The approach works by counting matching n-grams in the candidate translation to n-grams in the reference text, where 1-gram or unigram would be each token and a bigram comparison would be each word pair. The comparison is made regardless of word order.

BLEU score is computed using the below formula

$$BLEU(N) = BrevityPenalty * GeometricAveragePrecisionScores(N) \tag{1}$$

Where,

$$GeometricAveragePrecision(N) = \exp(\sum_{n=1}^{N} w_n log p_n) = \prod_{n=1}^{N} {p_n}^{w_n} \tag{2}$$

$$= {p_1}^{1/4} * {p_2}^{1/4} * {p_3}^{1/4} * {p_4}^{1/4} \tag{3}$$

$$BrevityPenalty = \begin{cases} 1, & \text{if } c > r. \\ \exp(1 - r/c), & \text{if } c <= r. \end{cases} \tag{4}$$

Here $N = 4$ and uniform weights $w_n = N/4$. where p1 = precision 1-gram, p2 = precision 2 gram, p3 = precision 3 gram and p4 = precision 4 gram. 'c' is predicted length = number of words in the predicted length and 'r' is target length = number of words in the target sentence.

4.2 $\text{BLEU}_{\text{normalized}}$

Instead of calculating the BLEU scores on the texts where the Hindi words are transliterated to Roman, we calculate the score on texts where Hindi words are in Devanagari and English words in Roman.

5 Results & Error Analysis

5.1 mBART model

5.1.1 Results on Validation Set

Model	BLEU Score	BLEU Normalized
mBART-en	14.3	17.7
mBART-hien	14.5	19.7

Table 4: Results

5.1.2 Epoch Vs Loss and Epoch Vs BLEU Score plot of mBARTen model

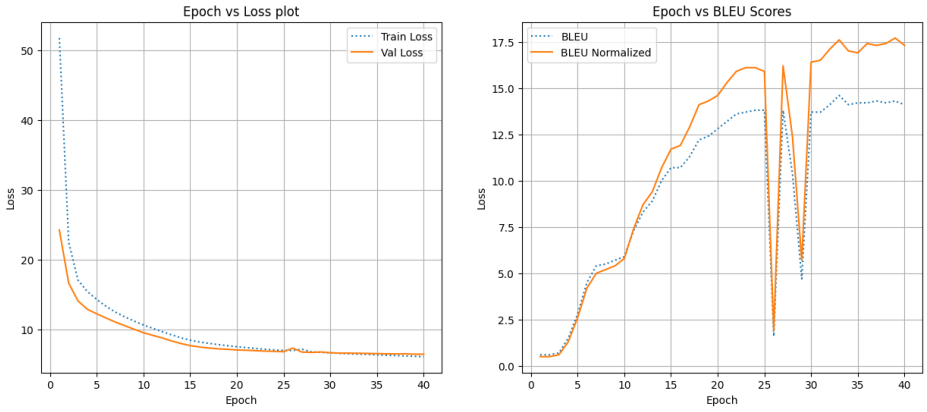

Figure 4: mBARTen results

5.1.3 Epoch Vs Loss and Epoch Vs BLEU Score plot of mBART-hien model

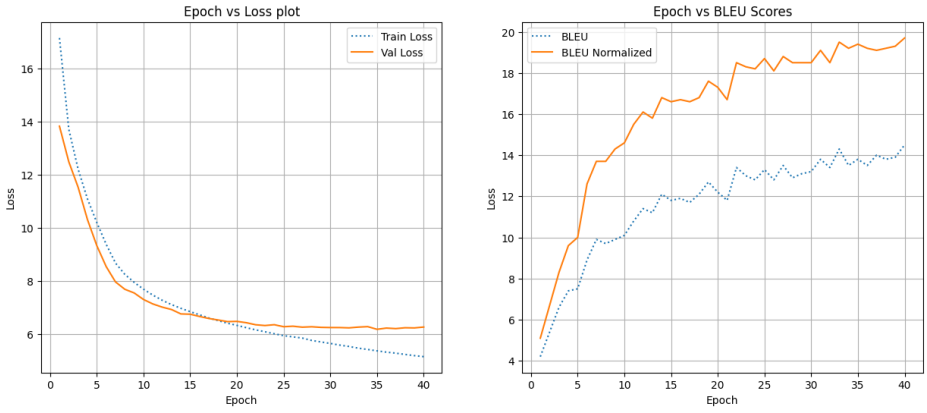

Figure 5: mBART-hien results

5.2 mT5 model

5.2.1 Results on Validation Set

Model	BLEU Score
mT5 English-Hindi Codemix	8.94
mT5 English-Telugu Codemix	3.818

Table 5: Results

5.2.2 Results Plot of mT5 English-Hindi Codemix Model

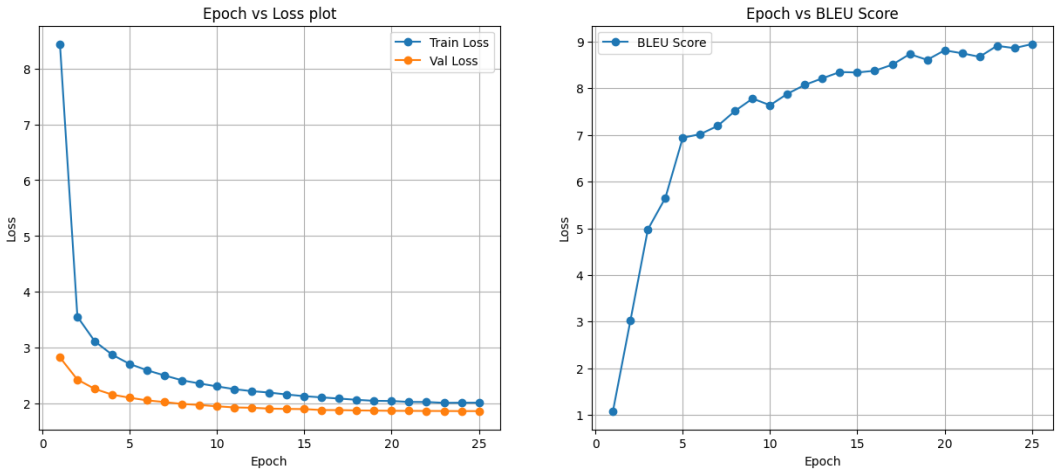

Figure 6: mT5 eng-hin results

5.2.3 Results Plot of mT5 English-Telugu Codemix Model

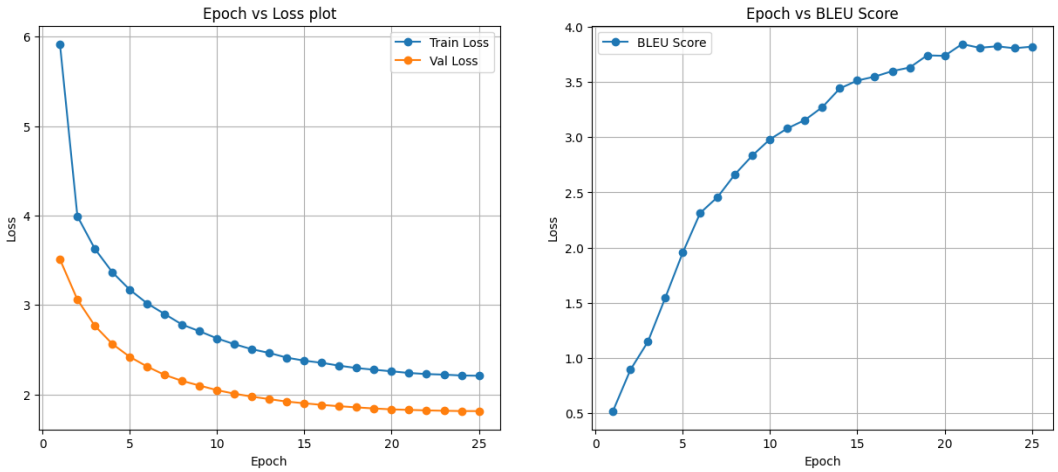

Figure 7: mT5 eng-tel results

5.3 IndicBART model

5.3.1 Results on Validation Set

Model	BLEU Score
mT5 English-Hindi Codemix	11.83
mT5 English-Telugu Codemix	0.518

Table 6: Results

5.3.2 Results Plot of IndicBART English-Hindi Codemix Model

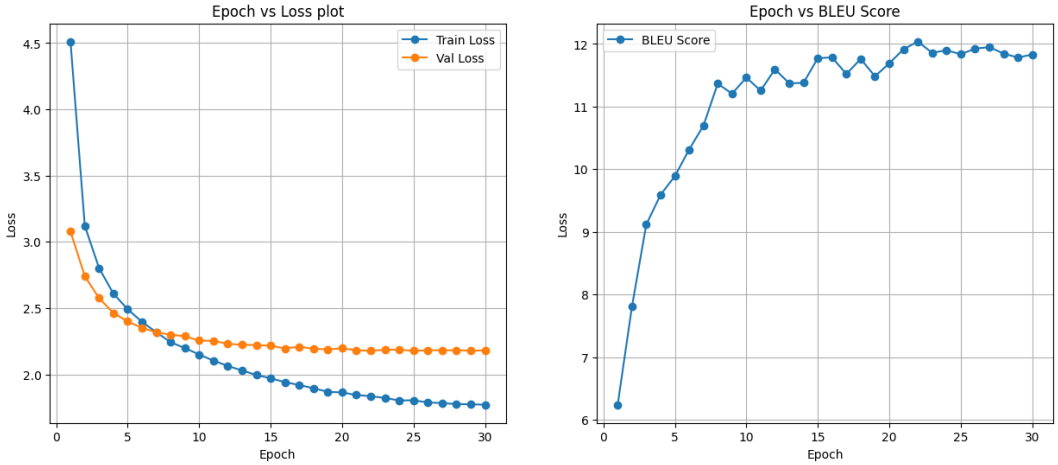

Figure 8: IndicBART eng-hin results

5.3.3 Results Plot of IndicBART English-Telugu Codemix Model

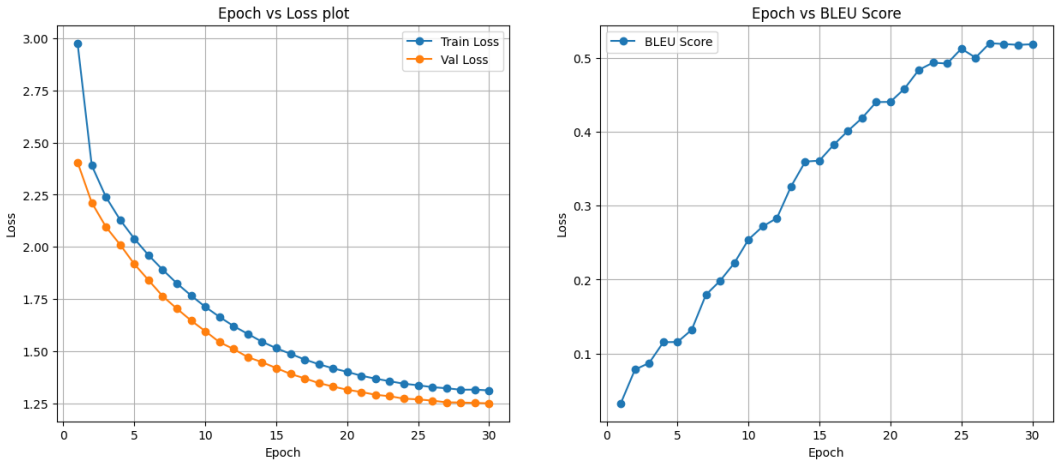

Figure 9: IndicBART eng-tel results

5.4 Analysis of Outputs

English Sentence (Input)	Generated Sentence (CodeMix)	Comment
I think they have a strong defense to beat the Chitauri invaders.	मुझे लगता है कि उनके पास Chitauri invaders को beat करने का strong defense है	Perfect translation
the ratings are very low for both rotten tomatoes and metacritic	rotten tomatoes और metacritic की ratings बहुत कम है	Perfect translation
I do. I adore sci-fi. You?	मैं करता हूँ. मैं sci-fi adore करता हूँ. आप?	Perfect translation
I think you watched this movie also	मुझे लगता है कि मैंने यह movie भी देखी है	meaning changed
Me too. The more unexpected, the better!	मैं भी. ज्यादा unexpected है	Incomplete or missing info
Yeah, even though I love comedies, this wasn't 1 of my faves	हां, even though मुझे comedies पसंद हैं लेकिन मेरे fades में यह नहीं है	Factually incorrect

Figure 10: English-Hindi CodeMix outputs analysis

English Sentence (Input)	Generated CodeMix (English-Telugu)	Comment
My voice in Telugu is very good.	తెలుగు లో నా వాయిస్ చాలా బాగుంది.	Perfect translation
Raj Tarun Power Play Movie Review Tell Bro	రాజ్ తరుణ్ పవర్ ప్లే మూవీ రివ్యూ చెప్పు బ్రో	Perfect translation
Hello Thyview. Waiting for Master Review	హెల్లో థైవ్యూ. మాస్టర్ రివ్యూ కోసం వెయిటింగ్	Perfect translation
Locations are the main plus points of the movie	మూవీ లోనే లాజిక్స్ మాత్రం మెయిన్ పాయింట్స్	meaning changed
Bro talk in normal language	బ్రో లాంగ్వేజ్ లో మాట్లాడు బ్రో	missing info
@Chandler999999 I saw Knight and it was great.	@కౌంటర్౯౯౯౯ నేను కోటిని చూసాను అది గ్రేట్ గా ఉంది.	Factually incorrect

Figure 11: English-Telugu CodeMix outputs analysis

These are some of the examples of English-Hindi and English-Telugu outputs generated by the models, On Hindi code mix test data the model has produced decent results, and most of them are almost near perfect translations when we manually analysed the outputs. On the other side the BLEU score is low, which proves the inability of the BLEU metric to capture the semantic meaning of the text. On Telugu code mix dataset the models failed to perform well, and we observe that the quality of the Telugu dataset is poor, because it is not annotated by humans, and we rely on tranliteration and translation to prepare English sentences, we end up getting a noisy dataset to train the models and it is evident from the results that model failed to perform well.

6 Conclusion & Future Work

We have implemented the given baseline research paper and reproduced the results. The model mBART-hien has produced better results on the vcalidation dataset with a BLEU score of 14.5 and a normalized BLEU score of 19.7. We plotted the loss and BLEU scores for each epoch obtained from the model. In addition to that, we have also experimented with mT5 and IndicBART and report the BLEU scores.

We also made an attempt to extend this work to multiple Indian languages, as a first attempt we experimented on Telugu-English dataset, in which we make use of code mix data used for sentiment analysis and converted that into machine translation dataset by transliterating and translating. We report the performance on this dataset along with output analysis.

7 References/Reading Materials

References

[1] Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, Ponnurangam Kumaraguru (2021). CoMeT: Towards Code-Mixed Translation Using Parallel Monolingual Sentences. In NAACL 2021, Computational Approaches to Linguistic Code Switching Workshop..

[2] Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. Enabling Code-Mixed Translation: Parallel Corpus Creation and MT Augmentation Approach. In Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pages 131–140, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[3] 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), 2022, pp. 684-688, doi: 10.1109/ICSCDS53736.2022.9760834.

[4] 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.

[5] Attri, S.H., Attri, S.H., Prasad, T. and Ramakrishna, G. 2020. HiPHET: A Hybrid Approach to Translate Code Mixed Language (Hinglish) to Pure Languages (Hindi and English). Computer Science. 21, 3 (Sep. 2020). DOI:https://doi.org/10.7494/csci.2020.21.3.3624.