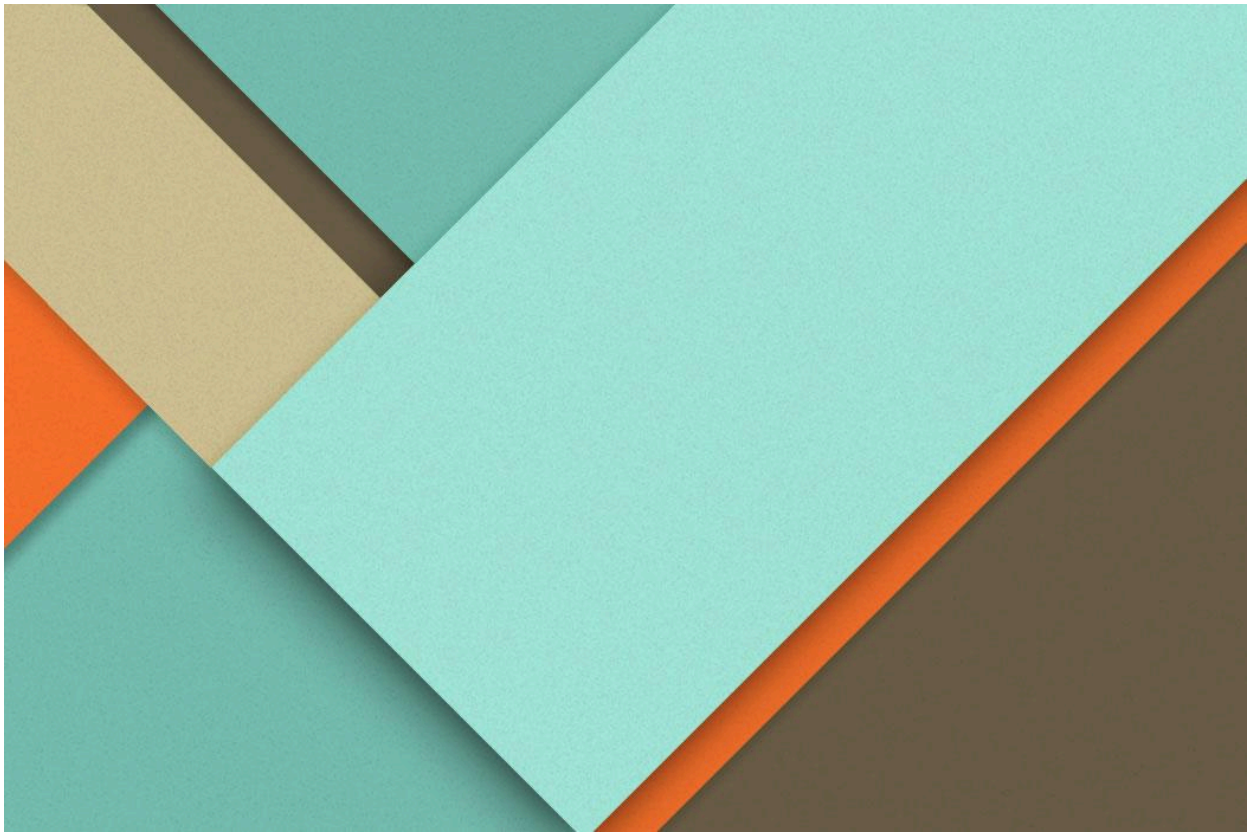

INLP Project Report

Code Mix Generation

Omprateek Shrivastava (2023201069)

Vishnu Ranjith (2023202004)

Pradhyumna Palore (2023202022)

Abstract

This project focuses on the generation of code-mixed text, a linguistic phenomenon observed in bilingual and multilingual societies where speakers seamlessly integrate multiple languages within a single sentence. Our primary objective is to develop algorithms that can generate code-mixed text, particularly in Hindi and English, emulating the linguistic fusion commonly observed in Hinglish. We employ a variety of techniques, including statistical methods and neural network architectures such as Long Short-Term Memory (LSTM), Transformer, and T5 models, to achieve this goal.

The evaluation of our code-mixed text generation models is performed using the BLEU (Bilingual Evaluation Understudy) score, a widely used metric in machine translation tasks. This report presents a detailed analysis of our experimental setup, methodologies employed, and the results obtained. Through systematic experimentation and evaluation, we aim to assess the effectiveness and performance of different approaches in generating code-mixed text. Insights gained from this research endeavor are expected to contribute to the advancement of natural language processing techniques and facilitate better understanding and modeling of code-mixing phenomena in multilingual communication contexts.

Objectives

1. Develop Algorithms: The primary objective of this project is to develop algorithms capable of generating code-mixed text. These algorithms should effectively integrate two or more languages, with a focus on Hindi and English, to mimic the natural code mixing observed in bilingual and multilingual societies.
2. Model Exploration: Explore a variety of techniques and architectures for code-mixed text generation, including statistical methods and neural network models such as Long Short-Term Memory (LSTM), Transformer, and T5. Investigate the strengths and limitations of each approach in capturing the intricacies of code mixing.
3. Evaluation Metrics: Employ the BLEU (Bilingual Evaluation Understudy) score as the primary evaluation metric to assess the quality and fluency of the generated code-mixed text. Compare the performance of different algorithms and models based on their BLEU scores to determine their effectiveness in generating linguistically accurate and contextually relevant code-mixed text.

Introduction

Code mixing, a linguistic phenomenon prevalent in bilingual and multilingual societies, involves seamlessly integrating two or more languages within a single sentence or discourse. This blending of languages, exemplified by Hinglish—a fusion of Hindi and English—reflects the cultural and linguistic hybridity of modern societies, particularly evident in digital communication platforms.

In this project, we aim to develop algorithms for generating code-mixed text, focusing on Hindi and English. Initially, we utilized the Hinglish dataset but encountered challenges due to its limited size and diversity. To address this, we adopted a data augmentation approach, merging various datasets to create a customized corpus rich in linguistic variations.

This report discusses our methodology, experiments, and findings regarding code-mixed text generation. We highlight challenges in dataset acquisition and preprocessing and explore the implications of our research for natural language processing, aiming to enhance NLP applications in multilingual contexts.

	en	fr
0	Press houndouts should be printed	[start] Press houndouts होने चाहिए । [end]
1	called for explanation or information	[start] explanation या information [end]
2	Undertaking inspection and analysis of importe...	[start] आयतित एवं indigenous उर्वरकों का inspe...
3	for the People of the Right Hand	[start] दाहिने हाथ में नामए आमाल लेने People क...
4	These things tend to happen every years	[start] यह परिवर्तन हर years में होता है [end]
...	...	...
248325	Disequilibrium create a situation where it is ...	[start] Disequilibrium की situation का निर्माण...
248326	Formal site for Sanskrit OCR	[start] Sanskrit OCR की वैकल्पिक साइट [end]
248327	Message of the day	[start] day का Message [end]
248328	Not so are the prayerful	[start] मगर जो लोग prayerful हैं [end]
248329	Between Rs Rs	[start] Rs से Rs तक [end]

248330 rows × 2 columns

Experiments and Results

In our experiments, we employed three distinct models for code-mixed text generation: Long Short-Term Memory (LSTM), Transformer, and T5. Each model offers unique advantages and mechanisms for processing sequential and contextual information, contributing to its efficacy in generating code-mixed text.

I. LSTM

- Model Description: LSTM is a type of recurrent neural network (RNN) designed to capture long-range dependencies in sequential data. Unlike traditional RNNs, LSTM networks utilize a memory cell and gating mechanisms (such as input, output, and forget gates) to control the flow of information, mitigating the vanishing gradient problem and enabling better retention of context over longer sequences.
- Functionality: LSTM processes input sequences step-by-step, updating its internal state at each time step based on the current input and the information stored in its memory cell. For code-mixed text generation, LSTM can effectively learn the mappings between English and Hinglish words, capturing linguistic patterns and generating contextually appropriate code-mixed output.
- The BLEU Score for LSTM was 1.746 * 1e-3.

```
BLEU Score: 0.001746703231174558
```

- Below are some Example translations.

```
It was a cold evening laden with smoke => यह cold एक smoke से उत्तेजित shade है ।
And Yunus was most surely of the apostles => और निश्चय ही रसूलों में से है कि वह अनक़रीब भी satisfied गया
A fund created to produce for depreciation => depreciation के लिए सृजित fund ।
 elephants used for the construction work  => elephants के construction के लिए प्रयोग के लिए elephants
MMR The Facts MMR => mmr
They looked on her behavior as childish => वे अपनी behavior पर उनके behavior को देखा
Our names are on top of the line of promotion => हमारे names promotion line के top पर हैं
So whoever wills shall remember it => तो जो wills उसे याद कर ले
Relative strengths and weaknesses of various organizational forms => various संगठनात्मक रूपों की शक्तियां तथा क
मज़ोरियां
He had sudden attack of idiopathic epilepsy => वह idiopathic epilepsy का sudden attack थी
```

II. Transformer

- Model Description: The Transformer architecture, introduced by Vaswani et al. (2017), revolutionized natural language processing by dispensing with recurrent connections in favor of self-attention mechanisms. Transformers leverage multi-head self-attention layers and position-wise feedforward networks to capture global dependencies in input sequences more efficiently.

- Functionality: In a Transformer model, self-attention mechanisms allow each word in the input sequence to attend to all other words, capturing contextual information and dependencies in parallel. This enables the model to process input sequences more effectively, facilitating better understanding and representation of language patterns. For code-mixed text generation, Transformer models excel at capturing the semantic relationships between English and Hinglish words, enabling accurate and fluent generation of code-mixed output.
- The BLEU Score for LSTM was 2.157 * 1e-3.

```
BLEU Score: 0.0021579560805122444
```

- Below are some Example translations.

  Input :-

```
This fact is based on possibility
In Stanford Roy met many academicians and political workers
A group of worms which are parasites of plants
and the answer is it depends
I borrow the one from this entire thirty
Licencing and import policies were liberalised
Hey asshole don t forget your tip
Early reading and writing
Focus on Previous Folder
Turn each middle screw and cast puck together
never run out It is simply too common
He plays World of Warcraft
But such occasions should be few and far between
The Congress should remove such differences not create them
The questions are of four types
```

  Output :-

```
fact की यह समझता है कि inevitable returns पर
himalayas में political और political workers
worms जो एक indian के पौधे जो पौधे पैदा होते हैं ।
और answer है यह है
मैं entire से मिल जाएगा
mega policies व incentives भी
ओ credit आपका धन्यवाद
older reading और writing पढ़ा जा रहा है
previous folder पर focus
हर middle और एक copyright बनी इसराइल को उपवास करे
कभी कभी यह common नहीं है ।
वह world वॉरक्राफ़्ट खेलता है ।
लेकिन such few और few लोगों के बीच भी पुरस्कृत हुई ।
congress के लिए such differences नहीं होनी चाहिए ।
questions के चार types हैं
```

III. T5 (Text-To-Text Transfer Transformer)

- Model Description: T5 is a variant of the Transformer architecture proposed by Raffel et al. (2019) that frames all NLP tasks as text-to-text transformations. By

formulating tasks in a unified text-to-text format, T5 simplifies model design and training, enabling it to excel in a wide range of natural language understanding and generation tasks.

- Functionality: T5 processes input data in a text-to-text format, where both the input and output are represented as text sequences. This unified framework allows T5 to handle diverse NLP tasks, including code-mixed text generation, by conditioning the model on input sequences and generating corresponding output sequences. For code-mixed text generation, T5 leverages its ability to learn complex text transformations, enabling it to seamlessly convert English text to Hinglish while preserving context and fluency.
- The BLEU Score for LSTM was 2.542 * 1e-3.

```
BLEU Score: 0.0025425103103432468
```

- Below are some Example translations.

  Input :-

```
This fact is based on possibility
In Stanford Roy met many academicians and political workers
A group of worms which are parasites of plants
and the answer is it depends
I borrow the one from this entire thirty
Licencing and import policies were liberalised
Hey asshole don t forget your tip
Early reading and writing
Focus on Previous Folder
Turn each middle screw and cast puck together
never run out It is simply too common
He plays World of Warcraft
But such occasions should be few and far between
The Congress should remove such differences not create them
The questions are of four types
```

  Output :-

```
यह fact possibility पर based है ।
Stanford Roy met many academicians और political workers
worms group जो plants के parasites हैं ।
और answer यह depends है
मैं इस entire thirty से एक एक ।
Licencing और import policies liberalised
वहhole don t forget t tip
Early reading और writing
Previous Folder पर Focus
Turn each middle screw और cast puck ।
मेरी यह है कि यह common तीny है
वह Warcraft का World
लेकिन such occasions few और far को ।
Congress such differences को h से नहीं कर सका
सकते types के questions हैं ।
```

Evaluation Metric

In our experiments, we utilized the BLEU (Bilingual Evaluation Understudy) score as the primary evaluation metric to assess the quality and fluency of the generated code-mixed text. BLEU is a widely used metric in machine translation tasks, measuring the similarity between the generated text and reference text based on n-gram precision.

The BLEU score is calculated based on the following components:

- Precision: The precision measures how many n-grams in the generated text match those in the reference text. It is computed as the ratio of the number of overlapping n-grams to the total number of n-grams in the generated text.
- Brevity Penalty: The brevity penalty penalizes shorter translations to discourage overly concise outputs compared to the reference. It is calculated as the exponent of 1 minus the ratio of the length of the reference text to the length of the generated text.
- Modified n-gram precision: To avoid penalizing longer n-grams too much, BLEU uses modified precision, which computes the precision of each n-gram separately and then averages them.

The final BLEU score is computed as the geometric mean of the modified precision scores, with the brevity penalty applied.

The BLEU scores obtained for each model are as follows:

- LSTM: BLEU Score = 0.001746703231174558
- Transformer: BLEU Score = 0.0021579560805122444
- T5: BLEU Score = 0.0025425103103432468

These BLEU scores represent the quality of the generated code-mixed text compared to the reference code-mixed text. While all models achieved relatively low BLEU scores, indicating room for improvement, the T5 model demonstrated the highest BLEU score among the three, suggesting that it generated code-mixed text that was slightly closer to the reference text compared to LSTM and Transformer models.

However, it's important to note that BLEU scores can vary depending on factors such as dataset size, diversity, and complexity, as well as the specific characteristics of the code-mixed text being generated. Thus, while BLEU scores provide a quantitative measure of performance, they should be interpreted alongside qualitative evaluations to gain a comprehensive understanding of model effectiveness and identify areas for improvement.

Problems we faced

- Dataset Limitations: One of the primary challenges encountered in this project was related to the limitations of available datasets for code-mixed text generation. Initially, reliance on the Hinglish dataset proved insufficient due to its limited size and lack of diversity. This hindered the robust training of our models and affected the quality of generated code-mixed text. Despite efforts to augment the dataset by merging multiple sources, ensuring sufficient data diversity remained a persistent challenge.
- Computational Resources: Another significant hurdle was the intensive computational requirements associated with training and evaluating neural network models for code-mixed text generation. As the size and complexity of the models increased, so did the demand for computational resources, including processing power, memory, and storage. This led to computational bottlenecks and extended training times, impeding the efficiency of our experimentation process.
- Memory Exceeding Issues: The memory requirements of neural network models, particularly Transformer and T5 architectures, posed significant challenges during training and inference phases. As the models grew in size and complexity, memory usage exceeded available resources, leading to crashes, out-of-memory errors, and system instability. Addressing these memory exceeding issues necessitated optimization strategies, including batch size adjustments, model pruning, and utilization of distributed computing resources.
- Data Preprocessing Complexity: Preprocessing code-mixed text data, particularly converting English text to Hinglish, introduced additional complexity and computational overhead. Implementing robust preprocessing pipelines to handle language conversion, tokenization, and data cleaning required considerable effort and expertise. Moreover, ensuring the preservation of linguistic and contextual nuances during preprocessing posed challenges, impacting the quality of the final dataset and subsequent model performance.

Addressing these challenges required a multidisciplinary approach encompassing expertise in natural language processing, machine learning, and computational resources management. Strategies such as data augmentation, optimization of model architectures, utilization of parallel processing techniques, and efficient memory management were essential to mitigate these challenges and advance the development of code-mixed text generation models. Despite these obstacles, ongoing research efforts aim to overcome these challenges and enhance the effectiveness of code-mixed text generation in diverse linguistic contexts.

Conclusion

In conclusion, our project aimed to tackle the challenging task of code-mixed text generation, focusing on the fusion of Hindi and English languages, exemplified by Hinglish. Through systematic experimentation and evaluation, we explored the effectiveness of three prominent models: Long Short-Term Memory (LSTM), Transformer, and T5, in generating code-mixed text. While our endeavor encountered several challenges, including dataset limitations, computational resource constraints, and memory exceeding issues, it provided valuable insights into the complexities of code mixing and the capabilities of state-of-the-art natural language processing models.

Despite the challenges faced, our experiments yielded promising results, with the T5 model demonstrating the highest BLEU score among the three models, indicating relatively better performance in generating code-mixed text closer to the reference. However, it is essential to interpret these results cautiously, considering the inherent limitations of BLEU scores and the need for qualitative assessments to complement quantitative evaluations.

Moving forward, further research and development efforts are warranted to address the challenges identified in this project and advance the state-of-the-art in code-mixed text generation. This includes expanding and diversifying datasets, optimizing model architectures for efficiency and scalability, and exploring innovative techniques to enhance the quality and fluency of generated code-mixed text. Additionally, collaboration across multidisciplinary domains, including linguistics, computational linguistics, and machine learning, is crucial to foster innovation and progress in this rapidly evolving field.

In summary, our project contributes to the growing body of knowledge in code-mixed text generation, paving the way for future advancements in multilingual natural language processing and facilitating better understanding and communication in diverse linguistic communities. By addressing the challenges and leveraging the opportunities presented in this research, we aim to empower individuals and organizations to navigate the complexities of code mixing and embrace linguistic diversity in the digital age.