Code Mixed Generation

Team 30

Introduction

Code mixing is the use of multiple languages in a single sentence, often seen in bilingual or multilingual communities. Code mix generation in Natural Language Processing (NLP) refers to the creation of code-mixed text using statistical methods or neural networks. Code mix generation has applications in fields such as machine translation, speech recognition, and sentiment analysis.

Project Proposal

- **Code Mixed Generation**:

  Generate code mixed sentences.

- **Code Mixed Translation**: Translate

  English to Code Mixed sentences.

Part I : Code Mixed Generation

Generating code mixed (Hinglish) text…

Tasks:

- Data Exploration
- Data Pre-processing
- Model Creations
- Hyper-Parameter Tuning
- Metric Analysis and reports

Models

1. **Baseline Model** : Single LSTM based model to predict next word
2. **Improved Model** : Two LSTM based model, one to encode language-id, another to predict the word.

Generated Sentences

Baseline

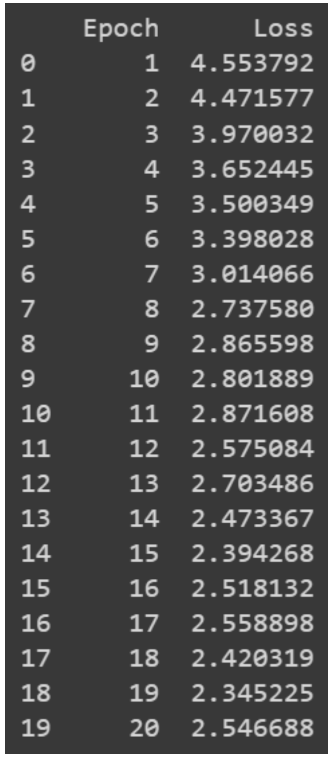
Seed	Generated
you	you can mein character ka unique hai to kya hum is
me abhi	me abhi bhi yanhi nahi hota achyar bhi toy story dekhi usme
hi me	hi me se he jo app over par can interesting sound kartha
me sochta	me sochta nahi hu but <number> ghante kahana hoga aur voh enigma
life me always	life me always aisa hi lagta tha wo karke jo usko mila wo

Improved

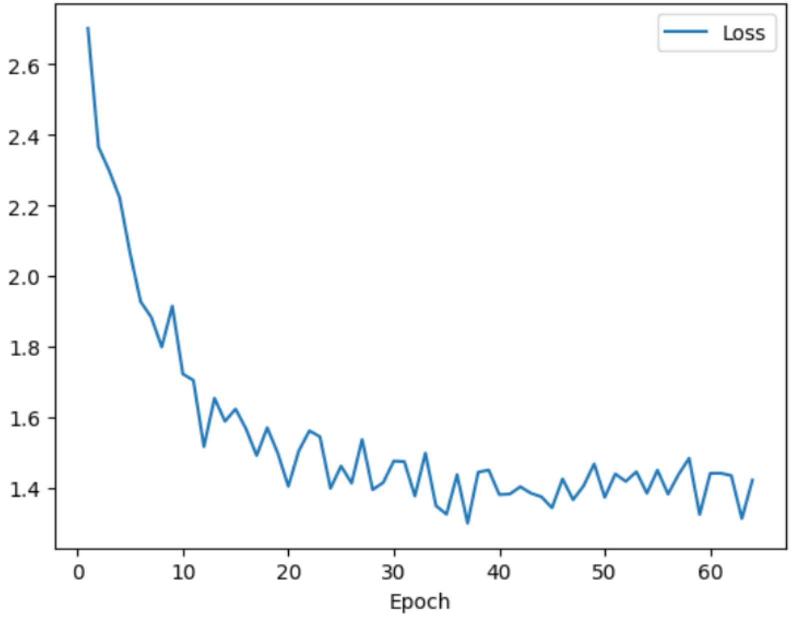
Seed	Generated
you	you seen waise mai use bar dekh sakta hun ki unpredictable
me abhi	me abhi tak same cheej he jitni kuch der pahale ye movie
hi me	hi me nahi dekhi main sochta hoon ke wo lucky ho raha
me sochta	me sochta hu wwe nahi karte even if they dont recall the
life me always	life me always wonder karti hai yah use itna trouble kiya <end> ne

Metrics and performance

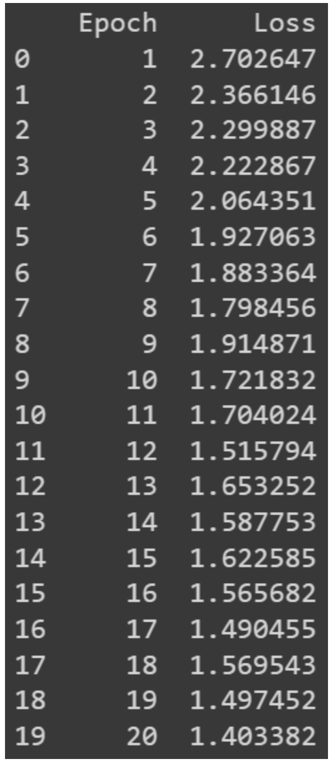
Baseline



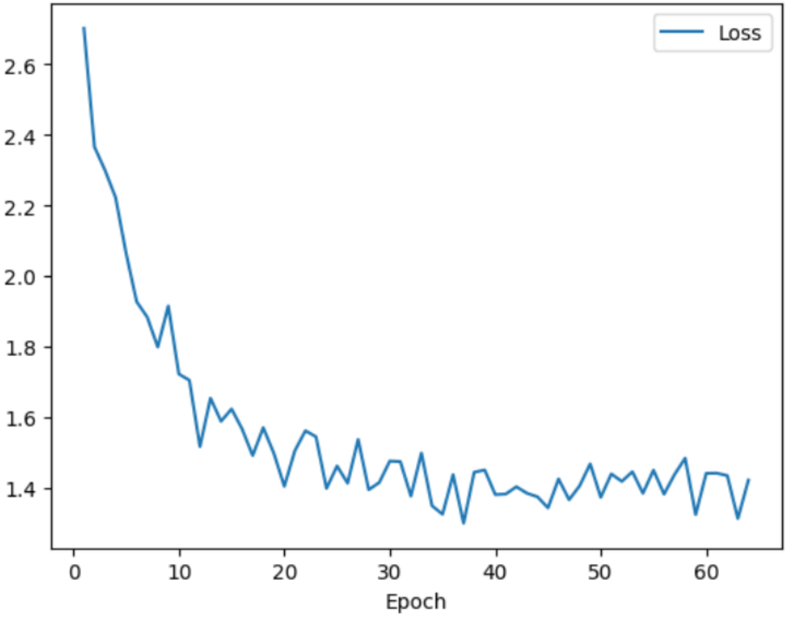
```
    Epoch      Loss
0       1  4.553792
1       2  4.471577
2       3  3.970032
3       4  3.652445
4       5  3.500349
5       6  3.398028
6       7  3.014066
7       8  2.737580
8       9  2.865598
9      10  2.801889
10     11  2.871608
11     12  2.575084
12     13  2.703486
13     14  2.473367
14     15  2.394268
15     16  2.518132
16     17  2.558898
17     18  2.420319
18     19  2.345225
19     20  2.546688
```



Improved



```
    Epoch      Loss
0       1  2.702647
1       2  2.366146
2       3  2.299887
3       4  2.222867
4       5  2.064351
5       6  1.927063
6       7  1.883364
7       8  1.798456
8       9  1.914871
9      10  1.721832
10     11  1.704024
11     12  1.515794
12     13  1.653252
13     14  1.587753
14     15  1.622585
15     16  1.565682
16     17  1.490455
17     18  1.569543
18     19  1.497452
19     20  1.403382
```



Perplexity

Train Dataset	2.6851668370395037
Validation Dataset	216.5501515989133

Perplexity

Train Dataset	2.4366996327589905
Validation Dataset	6.972607726896829

Mix Factor (MF): 44.86507936507939

Mix Factor (MF): 50.10277777777779

Part II : Code Mixed Translation

Translating English text to Code Mixed (Hinglish)

Tasks:

- Data Exploration
- Data Pre-processing
- Model Creation
- Hyper-Parameter Tuning
- Metric Analysis and reports

Models

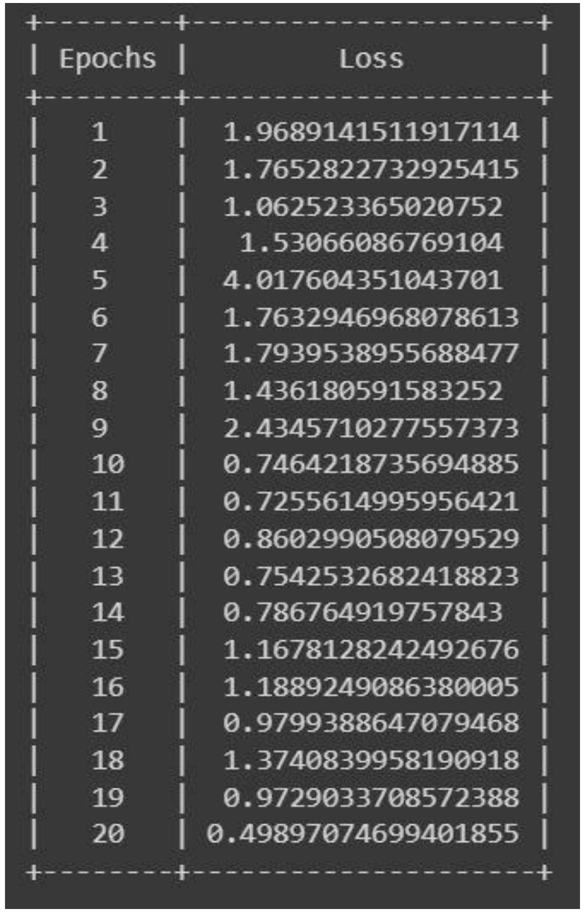
Baseline Model : Encoder Decoder Model using LSTM

Translated Sentences

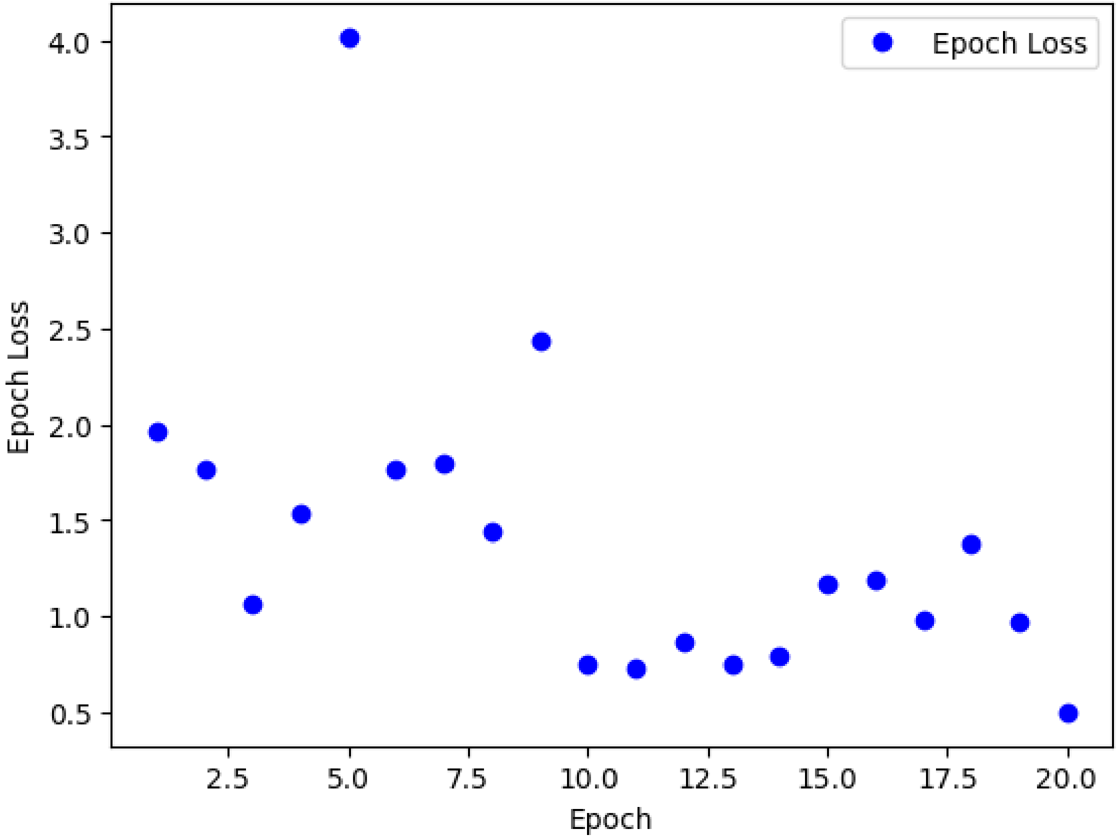
Baseline

English Sentence	Translated Sentence
Alright that is fine. What is the movie?	interesting hai kya ye ek long movie hai?
I have not seen that one either	maine kabhi nahi dekhi
may be worth watching!	do box bhi man!

Metrics and performance



```
+--------+---------------------+
| Epochs |         Loss        |
+--------+---------------------+
|   1    |  1.9689141511917114 |
|   2    |  1.7652822732925415 |
|   3    |  1.062523365020752  |
|   4    |   1.53066086769104  |
|   5    |  4.017604351043701  |
|   6    |  1.7632946968078613 |
|   7    |  1.7939538955688477 |
|   8    |  1.436180591583252  |
|   9    |  2.4345710277557373 |
|   10   |  0.7464218735694885 |
|   11   |  0.7255614995956421 |
|   12   |  0.8602990508079529 |
|   13   |  0.7542532682418823 |
|   14   |  0.786764919757843  |
|   15   |  1.1678128242492676 |
|   16   |  1.1889249086380005 |
|   17   |  0.9799388647079468 |
|   18   |  1.3740839958190918 |
|   19   |  0.9729033708572388 |
|   20   | 0.49897074699401855 |
+--------+---------------------+
```



BLEU SCORE: 1.38

Future Work

Generation

- Use Transformers.
- Using context based embeddings like BERT, ELMo

Translation

- Apply attention to Encoder – Decoder Models.
- Use Transformers and various pretrained models available.

References

- https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- https://medium.com/analytics-vidhya/machine-translation-encoder-decoder-model-7e4867377161
- https://towardsdatascience.com/perplexity-in-language-models-87a196019a94
- https://www.kdnuggets.com/2020/07/pytorch-lstm-text-generation-tutorial.html

Thank you

- Ashutosh gupta (2021201085)
- Sk Abukhoyer (2021201023)
- Soumodipta Bose(2021201086)