INLP PROJECT

MACHINE TRANSLATION

Omprateek Shrivastava
Vishnu Ranjith
Pradhyumna Palore

CONTENT

- TIMELINE
- CHALLENGES & LIMITAIONS
- DATASET USED
- LSTM MODEL
- TRANSFORMER
- LLM’S
- EVALUATION METRIC

TIMELINE

Outline Phase :- Analyzing initial architechture and exploration of Hinglish and related Datasets, identification of challenges.

Interim Phase :- Creation of a LSTM architechture, deciding which dataset would give better results, and noting down observations.

Final Phase - 1 :- Implementing different models and architechture, generating customized datasets and Hyperparameter tuning.

Final Phase - 2 :- Comparing models using BLEU Score, brainstorm and analyze on what works better, and what does not.

CHALLENGES AND LIMITATIONS

- **NUMBER OF EPOCHS / TIME** :- The training time on average was 1hr/epoch, which makes it impractical to train for larger number of iterations.
- **COMPUTE** :- The amount of compute to run a large dataset size on a model with given number of epochs is large.
- **LIMITATION OF DATASET** :- Datasets satisfying the requirements were difficult to find and limited leading to the creation of Custom dataset.

DATASET USED

- Initially used Hinglish Dataset with 10K sentences each of English and the corresponding Hinglish.
- Combined various datasets to make a combined dataset containing almost 2 lakh sentences.
- The Dataset are divided in the ratio 80:20.

	en	hing
0	Pension Fund Managers PFMs	Pension Fund PFMs
1	A new roll of blotting paper has been ordered	blotting paper के new roll का आदेश दिया गया है ।
2	Accelerated Irrigation Benefits Programme	त्वरित Irrigation Benefits Programme
3	So they go deep inside mines	इस लिए वह भूमि के अन्दर गहरे mines है
4	Right to constitutional remedies for enforceme...	Fundamental अधिकार के enforcement के लिए const...

LSTM

Parameters :

- Embedding Size : 300
- Hidden_size = 256
- num_epochs = 20
- lr = 0.001

Model Translation Example :

```
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
```

- LSTM performed poorly in terms capturing the inherent meaning of the sentence and translating the same into a hinglish sentence.
- It indicates that the sequence-wise analysis of the tokens is not sufficient to fulfill the objective.

TRANSFORMER

Parameters :

- Embedding Size : 256
- num_heads = 8
- num_epochs = 20
- lr = 0.001

Model Translation Example :

This fact is based on possibility	fact की यह समझता है कि inevitable returns पर
In Stanford Roy met many academicians and political workers	himalayas में political और political workers
A group of worms which are parasites of plants	worms जो एक indian के पौधे जो पौधे पैदा होते हैं ।
and the answer is it depends	और answer है यह है
I borrow the one from this entire thirty	मैं entire से मिल जाएगा
Licencing and import policies were liberalised	mega policies व incentives भी
Hey asshole don t forget your tip	ओ credit आपका धन्यवाद
Early reading and writing	older reading और writing पढ़ा जा रहा है
Focus on Previous Folder	previous folder पर focus
Turn each middle screw and cast puck together	हर middle और एक copyright बनी इसराइल को उपवास करे
never run out It is simply too common	कभी कभी यह common नहीं है ।
He plays World of Warcraft	वह world बॉरक्राफ़्ट खेलता है ।
But such occasions should be few and far between	लेकिन such few और few लोगों के बीच भी पुरस्कृत हुई ।
The Congress should remove such differences not create them	congress के लिए such differences नहीं होनी चाहिए ।
The questions are of four types	questions के चार types हैं

- Transformers were extremely efficient in terms of both time and resources.
- But it doesn't effectively capture the context of the sentence while translating.
- This leads to incoherent output.

LLM'S - T5

Parameters :

- Embedding Size : 512
- num_heads = 8
- num_epochs = 3
- lr = 0.001

Model Translation Example :

- T5 model worked extremely well in understanding the meaning of the sentence and was able to output a coherent and sensible Hinglish sentence in most cases barring few grammatical mistakes.

This fact is based on possibility	यह fact possibility पर based है ।
In Stanford Roy met many academicians and political workers	Stanford Roy met many academicians और political workers
A group of worms which are parasites of plants	worms group जो plants के parasites हैं ।
and the answer is it depends	और answer यह depends है
I borrow the one from this entire thirty	मैं इस entire thirty से एक एक ।
Licencing and import policies were liberalised	Licencing और import policies liberalised
Hey asshole don t forget your tip	वहhole don t forget t tip
Early reading and writing	Early reading और writing
Focus on Previous Folder	Previous Folder पर Focus
Turn each middle screw and cast puck together	Turn each middle screw और cast puck ।
never run out It is simply too common	मेरी यह है कि यह common तीny है
He plays World of Warcraft	वह Warcraft का World
But such occasions should be few and far between	लेकिन such occasions few और far को ।
The Congress should remove such differences not create them	Congress such differences को h से नहीं कर सका
The questions are of four types	सकते types के questions हैं ।

EVALUATION METRIC

BLEU SCORE

BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another.

BLEU Score for Various Models :

- LSTM : 1.746 * 1e-3
- Transformer : 2.157 * 1e-3
- T5 : 2.545 * 1e-3

GITHUB Link : *https://github.com/PradhyumnaPalore/HinglishGeneration*

THANK YOU