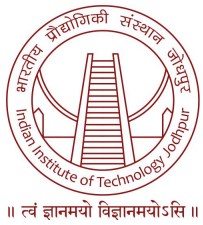

Translation from English to Indic Language

Objective: Implement a transformer-based encoder-decoder architecture for language translation from English to Indic Language

Dataset used: Samanantar, which is the largest corpora collection available for the Indic language.
Dataset link: https://indicnlp.ai4bharat.org/samanantar/
We have used English to Hindi language pairs, i.e. en-hi for training our model which consist of a total of 8.56M data.

Steps Involved:

1. Total count of English to Hindi dataset was 8466307. Due to computational issues, we have used only the first 200000 data for our analysis.
2. Train Test split of 80:20 was done for training and testing the model respectively.
3. Tokenized the input using *"MBart50TokenizerFast"* tokenizer.
4. Trained the MBart transformer model on our dataset.
5. Saved the trained model on our external memory.
6. The pretrained model was tested on the test dataset (first 200 entries) and different performance scores (*BLEU Score, ROUGE Score and METEOR Score)* were calculated corresponding to each predicted translation and ground truth for performance analysis.

Total Sentences	Training Sentences	Testing Sentences
200000	160000	40000

Training of model:

```
    model1(**model_inputs1, labels=labels1)  # forward pass
    if count==200000:
       print("Training done")
       print(model1(**model_inputs1, labels=labels1))
       model1.save_pretrained('/content/drive/MyDrive/My Documents/NLU/Project/model')
       break
```

```
Training done
Seq2SeqLMOutput(loss=tensor(2.9312, grad_fn=<NllLossBackward0>), logits=tensor([[[ -4.8904,  -4.7679,
         [ -1.6790,  -1.7942,   1.2779,  ..., -15.3255,  -7.0511,  -1.7737],
         [ -0.2676,  -0.3838,   2.7434,  ...,  -9.7866,  -5.8003,  -0.4227],
         ...,
         [  0.4922,   0.4425,   7.9551,  ...,  -3.3791,  -2.8182,   0.5734],
         [  0.3819,   0.2819,   9.7243,  ...,  -3.0061,  -1.1938,   0.2541],
         [  0.4769,   0.4702,  15.0996,  ...,   0.8606,   0.9420,   0.5002]]],
       grad_fn=<AddBackward0>), past_key_values=None, decoder_hidden_states=None, decoder_attentions=Nc
          -1.2677e-02,  1.9742e-03],
         [-7.1853e-01,  5.8236e-01, -4.4010e-01,  ...,  4.3614e-02,
          -3.6984e-01,  7.9844e-01],
         [-3.2674e-01,  1.7987e-01,  3.8536e-01,  ..., -5.6770e-01,
          -1.2726e-01,  6.0201e-01],
```

Evaluation

We have evaluated the performance of model on the basis of three different scores, which are:

- BLEU Score
- ROUGE Score

1. BLEU (Bilingual Evaluation Understudy Score) Score is used to evaluate machine translated text and ranges between the value 0 to 1. It uses similarity between the translated sentence and original sentence to calculate the score. It is basically a precision focussed metric which evaluates the n-gram overlap of reference and predicted sentences.

Interpretation of BLEU Score:

- 0 represents a perfect mismatch, indicating the translated sentence has no overlap with the original text are identical
- 0.3 - 0.4 : Understandable translation
- 0.4 - 0.6 : Good Quality Translation
- 0.6 above : Very High Quality Translation
- 1 represents a perfect match, indicating the translated and original text are identical.

2. ROUGE Score

ROUGE (Recall Oriented Understudy for Gisting Evaluation) score is similar to BLUE score, only difference is it is recall based instead of precision.

Model's Performance:

No. of test data	Avg. BLEU Score Computed for Eng-Hindi Translation
200	0.6393

No. of test data	Avg. ROUGE Score Computed for Eng-Hindi Translation
200	0.6139

As we can see that for,

- English-Hindi Translation, the BLEU Score is around 0.639 which is quite good. Analysis of few translated and original sentences showed that the context/gist of both the sentences were the same most of the time, only the manner in which they were said varied.

Architecture of the model used:

Sequence-to-Sequence Transformer Architecture with 12 encoder and decoder layers (each) were used in the model with a model dimension of 1024 on 16 heads. An additional layer normalization on top of encoder and decoder layers were used.

architectures	MBartForConditionalGeneration
model_type	mbart
transformers_version	4.18.0
d_model	1024
decoder_attention_heads	16
decoder_ffn_dim	4096
dropout	0.1
decoder_layers	12
encoder_attention_heads	16
encoder_ffn_dim	4096
encoder_layers	12
encoder_layerdrop	0.0

The model consists of Encoder and Decoder of 12 layers each.

Train Test Split:

Due to computational issues, we have used only the first 200000 data for our analysis. The train test split is in the ratio 80:20

<u>Examples of English-Hindi Translationsobtained from the model</u>

Sentence in English	Translation in Hindi
I enjoy doing my work and I give my best to it.	मैं अपने काम में आनंद लेता हूं और उसमें अपना पूरा प्रयास करता हूं।
This Annual Athletic Meet gave the students an opportunity to prove their sporting abilities and win laurels.	इस वार्षिक एथलेटिक्स मीट ने विद्यार्थियों को अपने खेल-कूद की क्षमता को सिद्ध करने और लारल जीतने का अवसर दिया।
The film will have Aamir Khan playing the titular role.	इस फिल्म में अमीर खान का शीर्षक भूमिका निभाया जाएगा।

This reflects the fact that in many programming languages these are the characters that may be used in identifiers.	यह इस तथ्य को प्रतिबिंबित करता है कि कई प्रोग्रामिंग भाषाओं में ये वे अक्षर हैं जिन्हें पहचानकर्ताओं में प्रयोग किया जा सकता है।

```
This Annual Athletic Meet gave the students an opportunity to prove their sporting abilities and win laurels.
Hindi Predicted Translation:  इस वार्षिक एथलेटिक्स मीट ने विद्यार्थियों को अपने खेल-कूद की क्षमता को सिद्ध करने और लारल जीतने का अवसर दिया।
Ground Truth:  इस वार्षिक खेल आयोजन ने छात्रों को अपनी खेल प्रतिभाओं को प्रदर्शित करने और पुरस्कार जीतने का अवसर प्रदान किया।
BLEU Score:  0.692035506755669
ROUGE Score: 0.569620253164557
METEOR Score: 0.446

**************************************************************************

I enjoy doing my work and I give my best to it.
Hindi Predicted Translation:  मैं अपने काम में आनंद लेता हूं और उसमें अपना पूरा प्रयास करता हूं।
Ground Truth:  मेहनत और लगन से काम करती हूं और अपने काम को एन्जॉय करती हूं।
BLEU Score:  0.7324912081306231
ROUGE Score: 0.5238095238095238
METEOR Score: 0.319

**************************************************************************

"""All evidence has been collected."
Hindi Predicted Translation:  "सभी साक्ष्य एकत्र किया गया है।"
Ground Truth:  सारे साक्ष्य एकत्र कर लिए गए हैं।
BLEU Score:  0.8274377299117183
ROUGE Score: 0.6333333333333333
METEOR Score: 0.264
```

Almost all the hindi translations obtained were correct conveying the required meaning. Though the way of speaking was a bit different.

References:

[1]https://neptune.ai/blog/hugging-face-pre-trained-models-find-the-best

[2]https://cloud.google.com/translate/automl/docs/evaluate#:~:text=BLEU%20(BiLingual%20Evaluation%20Understudy)%20is,of%20high%20quality%20reference%20translations.

[3] https://towardsdatascience.com/the-most-common-evaluation-metrics-in-nlp-ced6a763ac8b

[4]https://towardsdatascience.com/how-to-evaluate-text-generation-models-metrics-for-automatic-evaluation-of-nlp-models-e1c251b04ec1

[5]https://aclanthology.org/2021.naacl-main.312.pdf