Name:- Vaibhav Sheth Roll no. 231110054

Observations :-

While translating I noticed that it converts sentence form into numerical form.

For example in test.en file, line no. 1595 it's mentioned that it is 64 thousand but while translating from Gujarati to English with NLLB (translated_GtoE_NLLB.en) it translated it as 64000.

Same goes with line no. 843 in both these files.

So translating it does not represent true information of original sentence.

As words are converted into numerical or vice-versa there may be possibility that it does not match while calculating BLEU score and Rouge Score.

So because of this, there might be reduction in metrics values.

BLEU Score is better for Indic-Trans compared to NLLB for all 4 translations.

For Rouge Score, it is better for some translation in NLLB and for some translation in Indic-Trans.

Also, ChatGPT does not give Gujarati translations in first time. We have to tell it many times.

Learning from this Assignment :-

Learned about how to translate Indian languages using different models. NLLB translates sentences one by one while Indic-Trans translates multiple sentences into one batch as specified.

Also learned about how to evaluate performance of this different translation models. Especially calculation of Rouge Scores and BLEU Score.

Evaluation Metrics

Rouge Score:-

ROUGE a set of metrics commonly used in natural language processing and machine translation to evaluate the quality of text summaries or translations.

1. ROUGE-1 (Unigram Overlap):

- ROUGE-1 measures the unigram (individual word) overlap between the reference and system-generated text.

2. ROUGE-2 (Bigram Overlap):

- ROUGE-2 measures the bigram (sequence of two adjacent words) overlap between the reference and system-generated text.

3. ROUGE-L (Longest Common Subsequence):

- ROUGE-L measures the longest common subsequence (LCS) overlap between the reference and system-generated text.

BLEU Score :-

The BLEU (Bilingual Evaluation Understudy) score is a metric used to evaluate the quality of machine-generated translations or text summarizations by comparing them against one or more reference (human-generated) translations or summaries.

Comparing the performance of NLLB-200, Indic-Trans, and ChatGPT will showcase the strengths and weaknesses of different machine translation approaches. Understanding which model excels in specific language pairs and contexts can guide future model selection for translation tasks.