OverView Of Models

NLLB-200 with 600M parameters

- NLLB-200 is a machine translation model primarily intended for research in machine translation, - especially for low-resource languages. It allows for single sentence translation among 200 languages.

- Primary intended users: Primary users are researchers and machine translation research community

- NLLB-200 is trained on general domain text data and is not intended to be used with domain specific texts, such as medical domain or legal domain. The **model is not intended to be used for document translation.** The model was trained with input lengths **not exceeding 512 tokens**, therefore translating longer sequences might result in quality degradation. NLLB-200 translations can not be used as certified translations.

- In addition, the supported languages may have variations that our model is not capturing. Users should make appropriate assessments.

ai4bharat/indictrans2

- The first open-source transformer-based multilingual NMT model that supports high-quality translations across all the 22 scheduled Indic languages

- including multiple scripts for low-resouce languages like Kashmiri, Manipuri and Sindhi. Overall, the model supports five scripts Perso-Arabic (Kashmiri, Sindhi, Urdu), Ol Chiki (Santali), Meitei (Manipuri), Latin (English), and Devanagari (used for all the remaining languages).

What Is BLEU and ROUGE Score

BLEU Score

- BLEU (Bilingual Evaluation Understudy) is a metric used for evaluating the quality of machine-translated text, typically in the context of natural language processing and machine translation tasks. it is checking based on n-gram overlap.

- The BLEU score ranges from 0 to 1, where a higher score indicates a better translation. It is calculated based on the precision of n-grams (contiguous sequences of words) in the candidate translation compared to the reference translation(s). The precision is modified to avoid penalizing translations that are shorter than the reference translation(s).

- it is quick and inexpensive to calculate.
- It is easy to understand.
- It is language independent.
- It correlates highly with human evaluation.
- It has been widely adopted.

- The approach works by counting matching n-grams in the candidate translation to n-grams in the reference text, where 1-gram or unigram would be each token and a bigram comparison would be each word pair. The comparison is made regardless of word order.

- I used here Corpus BLEU Score

- NLTK also provides a function called corpus_bleu() for calculating the BLEU score for multiple sentences such as a paragraph or a document.

ROUGE Score

- ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially a set of metrics for evaluating automatic summarization of texts as well as machine translations

- it works by comparing an automatically produced translation against a set of reference transalations (typically model vs test data).

- To get a good quantitative value, we can actually compute the precision and recall using the overlap.

- imply put, recall (in the context of ROUGE) refers to how much of the reference summary the system summary is recovering or capturing.

- i used here below ROUGE

- ROUGE-N — measures unigram, bigram, trigram and higher order n-gram overlap
- ROUGE-L — measures longest matching sequence of words using LCS. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length.

- ROUGE-1 refers to overlap of unigrams between the system summary and reference summary. ROUGE-2 refers to the overlap of bigrams between the system and reference summaries.

Result Of ROUGE & BLEU by comapring model vs test data

	Rouge-1(Recall)	Rouge-1 (Precision)	Rouge-1 (F-Score)	Rouge-2(Recall)	Rouge-2 (Precision)	Rouge-2 (F-Score)	Rouge-l (Recall)	Rouge-l (Precision)	Rouge-l (F-Score)	BLEU Score
NLLB (en to gu)	0.466	0.486	0.472	0.220	0.228	0.223	0.439	0.458	0.4457	0.6052
NLLB (gu to en)	0.579	0.581	0.575	0.344	0.341	0.340	0.541	0.543	0.538	0.6796
NLLB (gu to hi)	0.6022	0.6145	0.6047	0.3736	0.3787	0.3738	0.5710	0.5827	0.5734	0.6750
NLLB (hi to gu)	0.5239	0.5366	0.5266	0.2821	0.2877	0.2830	0.5046	0.5161	0.5071	0.6488
IndicTrans(en to gu)	0.5069	0.5133	0.5068	0.2619	0.2639	0.2611	0.4817	0.4877	0.4816	0.6394
IndicTrans(gu to en)	0.6274	0.6250	0.6224	0.4112	0.4080	0.4067	0.5931	0.5906	0.5883	0.7226
IndicTrans(gu to hi)	0.6063	0.6068	0.6027	0.3749	0.3727	0.3713	0.5730	0.5737	0.5697	0.6784
IndicTrans(hi to gu)	0.5134	0.5214	0.5140	0.2670	0.2696	0.2665	0.4904	0.4978	0.4909	0.6421
ChatGPT(en to gu)	0.2922	0.3038	0.2945	0.1034	0.1053	0.1030	0.2824	0.2931	0.2844	0.4616
ChatGPT(gu to en)	0.7042	0.7193	0.7087	0.5295	0.5415	0.5333	0.6727	0.6884	0.6777	0.7581
ChatGPT(gu to hi)	0.3797	0.3847	0.3800	0.1899	0.1906	0.1896	0.3661	0.3700	0.3660	0.5374
ChatGPT(mr to hi)	0.3619	0.3644	0.3599	0.1208	0.1206	0.1196	0.3536	0.3561	0.3517	0.5481

Analysis From Evalution

possible reasons for the variation in scores:

Quality of Translation: The quality of the translation itself plays a significant role. A better translation is likely to have higher Rouge and BLEU scores.

N-gram Overlap: Rouge-1 and Rouge-2 scores depend on the overlap of unigrams (single words) and bigrams (pairs of words) between the candidate and reference translations. If the translations use similar phrases or sentence structures as the reference, the scores will be higher.

Sentence Length: Longer sentences may have lower Rouge and BLEU scores due to a lower chance of exact n-gram matches.

Synonym Usage: Different word choices can affect the Rouge and BLEU scores. If the candidate translation uses synonyms not present in the reference, it may have lower scores.

Grammar and Syntax: Translations that maintain proper grammar and syntax of the target language are more likely to receive higher scores.

Evaluation Methodology: Different evaluation methodologies and implementations of Rouge and BLEU can lead to slight variations in scores.

Data Variation: The nature of the dataset and the diversity of topics and languages can also impact the scores. Some translations may be inherently easier or harder to evaluate.

Model Subjectivity: Evaluating translations is inherently subjective. Different model may give slightly different scores based on their interpretation of the quality of the translations.

NLLB model demonstrates strong performance, especially for Gujarati to English translation, achieving high ROUGE scores across all metrics. This indicates that NLLB produces accurate and fluent translations from Guajrati to English. same for Indic Trans Also giving good result for Gujarati to English same for chatgpt also .

for Other Language Translation overall performance of IndicTrans is good compare to NLLB and chatGPT also giving good result for indic to english and indic to indic language , as IndicTrans trained on 22 indic language so it is giving good result and identify correct word with correct ranslation and also understanding context and other gramatical entity while transalting from one language to other .

chatgpt also giving decent result but giving less score because not understanding gender , grammar and contex of sentence