CS689: Computational Linguistics

ASSIGNMENT 3 REPORT

Aditya Katare (231110005)

Translation Direction	BLEU Score	ROUGE-1 Recall	ROUGE-1 Precision	ROUGE-1 F1 Score	ROUGE-2 Recall
EN to HI (ChatGPT)	0.7009	0.6230	0.6254	0.6209	0.3736
HI to EN (ChatGPT)	0.6415	0.5606	0.5503	0.5502	0.3039
GU to HI (ChatGPT)	0.7011	0.6453	0.6426	0.6418	0.4111
EN to HI (IndicTrans)	0.6821	0.6030	0.6154	0.6051	0.3394
HI to EN (IndicTrans)	0.7328	0.6397	0.6267	0.6294	0.3964
GU to HI (IndicTrans)	0.6374	0.5821	0.5761	0.5747	0.3269
EN to HI (NLLB)	0.6065	0.5706	0.6144	0.5863	0.3025
HI to EN (NLLB)	0.6439	0.5856	0.5987	0.5857	0.3583
GU to HI (NLLB)	0.5257	0.4429	0.4972	0.4639	0.1782

Table 1: BLEU and ROUGE scores for different translation directions (Part 1).

Translation Direction	ROUGE-2 Precision	ROUGE-2 F1 Score	ROUGE-L Recall	ROUGE-L Precision	ROUGE-L F1 Score
EN to HI (ChatGPT)	0.3767	0.3726	0.5886	0.5909	0.5867
HI to EN (ChatGPT)	0.2946	0.2955	0.5288	0.5205	0.5196
GU to HI (ChatGPT)	0.4125	0.4097	0.6189	0.6174	0.6161
EN to HI (IndicTrans)	0.3466	0.3406	0.5687	0.5805	0.5709
HI to EN (IndicTrans)	0.3831	0.3864	0.6059	0.5950	0.5968
GU to HI (IndicTrans)	0.3289	0.3250	0.5534	0.5479	0.5464
EN to HI (NLLB)	0.3388	0.3163	0.5325	0.5742	0.5477
HI to EN (NLLB)	0.3541	0.3520	0.5534	0.5645	0.5529
GU to HI (NLLB)	0.2034	0.1874	0.4139	0.4629	0.4327

Table 2: BLEU and ROUGE scores for different translation directions (Part 2).

Learning from Translation Model Performance

Model Performance: Across all translation directions, the NLLB model consistently yields lower ROUGE scores compared to ChatGPT and IndicTrans. This observation suggests that the NLLB model might struggle more in capturing the subtleties and nuances of translation tasks.

Directional Variation: Different translation models exhibit varying performance depending on the translation direction. For instance, when translating from English to Hindi, ChatGPT outperforms other models, whereas in the Hindi to English direction, IndicTrans shows relatively better performance.

Evaluation Metric: ROUGE scores serve as valuable metrics for assessing translation quality, focusing on n-gram overlap between generated translations and reference texts. Although all models achieve respectable ROUGE scores, there remains room for improvement, particularly in terms of capturing semantic equivalence and contextual understanding.

Language Pair Impact: The performance of translation models varies significantly across different language pairs (e.g., English-Hindi, Hindi-English, Gujarati-Hindi). This underscores the importance of considering language-specific challenges and linguistic nuances inherent in each translation task.

Potential Areas for Improvement: To enhance translation quality, models could concentrate on improving ROUGE-2 and ROUGE-L scores, which evaluate bigram overlap and long-range dependencies, respectively. Achieving this improvement may involve refining contextual understanding and ensuring better preservation of semantic coherence throughout the translation process. article geometry

Conclusion

The NLLB model doesn't do as well as ChatGPT and IndicTrans, which means it struggles with understanding languages deeply. Different translation directions need different approaches. While the models get okay scores, they can do better in understanding what words really mean. Different languages have different challenges. To make translations better, we should focus on understanding how words go together and how they make sense. So, we need to work on improving how well the models understand sentences.