CS 689A : Computational Linguistics

Assignment 3 - Sanjeev Kumar (231110044)

Question 3 : Learning from Assignment

We performed the comparative performance for all three below translation models as asked in assignment.

- NLLB - 200
- IndicTrans2
- ChatGPT

Dataset Used:

- Dataset provided in assignment pdf.
- For testing First two models we used 1000 random sentences from ***test*** file in folder ***wat2021-devtest***
- For testing of ChatGPT model, we used 50 random sentences and manually passed them using a prompt to translate in the desired language.

Languages Used:

- English
- Hindi
- Gujarati

Translation Direction	ROUGE-1 (Recall)	ROUGE-1 (Precision)	ROUGE-1 (F1)	ROUGE-2 (Recall)	ROUGE-2 (Precision)	ROUGE-2 (F1)	ROUGE-L (Recall)	ROUGE-L (Precision)	ROUGE-L (F1)	BLEU Score
Eng to Hindi (NLLB-200)	0.560	0.603	0.575	0.326	0.352	0.335	0.520	0.561	0.535	0.619
Hindi to Eng (NLLB-200)	0.931	0.961	0.943	0.895	0.927	0.908	0.930	0.960	0.942	0.912
Hindi to Gujarati (NLLB-200)	0.414	0.470	0.435	0.184	0.211	0.194	0.387	0.439	0.407	0.521
Gujarati to Hindi (NLLB-200)	0.508	0.535	0.515	0.292	0.309	0.297	0.472	0.496	0.479	0.570
Englisg to Hindi (IndicTrans2)	0.616	0.622	0.616	0.381	0.385	0.380	0.579	0.585	0.579	0.692
Hindi to English (IndicTrans2)	0.669	0.666	0.664	0.458	0.453	0.452	0.637	0.634	0.632	0.754
Hindi to Gujarati (IndicTrans2)	0.512	0.519	0.512	0.263	0.266	0.263	0.488	0.496	0.488	0.641
Gujarati to Hindi (IndicTrans2)	0.596	0.592	0.590	0.361	0.359	0.357	0.560	0.557	0.555	0.675
English to Hindi (ChatGPT)	0.534	0.538	0.533	0.235	0.237	0.235	0.487	0.491	0.486	0.599
Hindi to English (ChatGPT)	0.776	0.752	0.761	0.568	0.555	0.558	0.750	0.727	0.735	0.801
Hindi to Gujarati (ChatGPT)	0.351	0.355	0.351	0.115	0.116	0.116	0.329	0.332	0.329	0.513
Gujarati to Hindi (ChatGPT)	0.489	0.519	0.498	0.220	0.232	0.224	0.447	0.476	0.457	0.548

Table-1 : Rouge and BLEU scores for different translation directions of all three models mentioned above

Analysis of the scores in the table and inference:

- Eng to Hindi Translation:
    - NLLB-200: Achieves moderate scores with ROUGE-1 F1 score of 0.575 and BLEU score of 0.619.
    - IndicTrans2: Shows improvement over NLLB-200 with higher ROUGE and BLEU scores.
    - ChatGPT: Performs slightly lower than NLLB-200 with similar ROUGE-1 F1 score but lower BLEU score.
- Hindi to Eng Translation:
    - NLLB-200: Shows strong performance with high ROUGE and BLEU scores.
    - IndicTrans2: Maintains high scores, indicating consistent performance.
    - ChatGPT: Shows comparable performance with NLLB-200 and IndicTrans2, with slightly lower ROUGE-1 F1 score but comparable BLEU score.
- Hindi to Gujarati Translation:
    - NLLB-200: Achieves moderate scores, indicating decent performance.
    - IndicTrans2: Shows improvement over NLLB-200 with higher ROUGE and BLEU scores.
    - ChatGPT: Performs competitively with NLLB-200 and IndicTrans2, with similar ROUGE and BLEU scores.
- Gujrati to Hindi Translation:
    - NLLB-200: Achieves moderate scores, indicating decent performance.
    - IndicTrans2: Shows improvement over NLLB-200 with higher ROUGE and BLEU scores.
    - ChatGPT: Performs competitively with NLLB-200 and IndicTrans2, with similar ROUGE and BLEU scores.

Overall, IndicTrans2 consistently outperforms NLLB-200 across all translation directions, indicating its superiority in translation quality. ChatGPT shows competitive performance with NLLB-200 and IndicTrans2, suggesting its effectiveness as a neural machine translation model.

However, this performance can not reflect the true nature of the translation models due to the low test data. Based on the complexity of words and semantics in randomly chosen sentences for this task, some model may suffer compared to others in some cases while the same can outperform in other cases.

Reasoning behind these observations can be attributed to below parameters:

- **Model Architecture and Training Data:** Each translation system employs a different model architecture and may have been trained on different datasets. The architecture and the quality and diversity of training data significantly impact the model's ability to generalize and produce accurate translations.
- **Language Pair Complexity:** Some language pairs may inherently pose more challenges for translation due to differences in grammar, syntax, vocabulary, and cultural nuances. The complexity of the Eng to Hindi, Hindi to Eng, Hindi to Gujarati, and Gujarati to Hindi language pairs may vary, affecting the performance of the translation systems.
- **Preprocessing and Tokenization:** Differences in preprocessing techniques and tokenization methods can affect how effectively the models handle input text. Variations in tokenization granularity or handling of special characters can impact translation quality.
- **Evaluation Metrics:** Differences in the evaluation metrics used to assess translation quality (e.g., ROUGE and BLEU scores) can also contribute to variations in reported performance. Some metrics may favor certain types of translations or penalize others differently.
- **Resource Availability:** The availability of computational resources, including hardware infrastructure and access to large-scale training data, can impact the quality and robustness of the trained models.

Conclusion:

While it is clearly visible that indicTrans outperforms other models in Indic to Indic direction translation. Comparative analysis of NLLB and ChatGPT can not be concluded in terms of superiority. ChatGPT might have suffered due to the prompt quality and also the size of the test data. Since IndicTrans and NLLB are checked on 1000 random sentences from given dataset whereas ChatGPT is evaluated on 50 sentences only so it might not capture the true nature of the model.