CS772 – Assignment 2

Transliteration

Due 11th October 2025

What is Transliteration?

- Converting text from one input script to another
- It is a character level task
- It can be modeled as a sequence-to-sequence problem

“ मैं घर जा रहा हूँ” *“Mein ghar ja rha hoon”*

Your Assignment – due 11 Oct 25

- Download transliteration data (Aksharantar) from AI4Bharat HuggingFace page
  - https://huggingface.co/datasets/ai4bharat/Aksharantar/tree/main
  - Language: Hindi – Roman to Devanagari
  - There is quite a lot of data, you can sub-sample training data as per your available compute. Subsample smartly, but **limit the training data to 100k examples**.
  - Use the entire test set from the Aksharantar copus. The test set must not be used as a part of training data or for tuning the model. Use the validation set from Aksharantar or create your own.
- Train transliteration models using encoder-decoder models using:
  - LSTM
  - Transformer
  - **Use maximum of 2 layers**
- **LLM**: Prompting off-the-shelf models – proprietary or open-source.
  - For open-source models, you can use services like FireBase or DeepInfra that host these models.
- You can use pre-written/available/LLM-generated code, but you will have to explain what the code is doing.
  - Take this opportunity to understand how the theory translates to code
- Later you will be asked to implement some variant which will require you to modify the code
  - Replace the standard attention by local attention in transformer model

Assignment discussion

Template for

You have to strictly follow this format

Define Transliteration

- Input
- Output

Data downloading and cleaning

- How much of data did you use?
- What kind of sampling of the data did you do?
- From what source?
- Did you use any cleaning? If so, what and how and why?

LSTM Based Transliteration

- What did you read for this part of the assignment?
- Compare greedy decoding vs beam search
- Is your program running?
- If yes, give the GUI-based demo

Transformer Based Transliteration

- What did you read for this part of the assignment?
- Compare greedy decoding vs beam search
- Is your program running?
- If yes, give the GUI-based demo

LLM Based Transliteration

- What did you read for this part of the assignment?
- Which LLM did you use?
- Try with various values for temperature and top_p.
- Is your program running?
- If yes, give the GUI-based demo

Compare and contrast

- Give a tabular comparison of word-level exact accuracy and character-level F1
    - More about these metrics here: https://aclanthology.org/W15-3902.pdf (Section 3.1 and 3.2)
    - You can use this evaluation script: https://github.com/snukky/news-translit-nmt/blob/master/tools/news_evaluation.py

- Analyze and explain the observations
- What are the characters/character sequences that are difficult to transliterate correctly? Why?