CS772 – Assignment 2: Transliteration

Siddhartha Rajeev, Shanttanu Oberoi

IIT Bombay

October 18, 2025

What is Transliteration?

- Converting text from one script to another (e.g., Devanagari to Roman)
- A character-level sequence-to-sequence task
- Example:


```
{"unique_identifier": "hin1", "native word": "मैट्रोलॉजिस्ट", "english word": "maitrologist", "source": "AK-Freq"}
{"unique_identifier": "hin2", "native word": "पीएचडब्ल्यूसीएस", "english word": "phwcs", "source": "AK-Freq"}
{"unique_identifier": "hin3", "native word": "प्रतिद्वन्दियों", "english word": "pratidwandiyon", "source": "AK-Freq"}
{"unique_identifier": "hin4", "native word": "प्रतियुक्ति", "english word": "pratiyukti", "source": "AK-Freq"}
{"unique_identifier": "hin5", "native word": "एक्सिसटेंस", "english word": "eksisatens", "source": "AK-Freq"}
```


- Can be modeled using encoder-decoder architectures

Define Transliteration

- **Input:** Hindi words in Roman script
- **Output:** Equivalent Devanagari transliteration
- **Task:** Learn mapping between corresponding characters/sequences (not of the same length!)

Data Downloading and Cleaning

- Dataset: **Aksharantar (AI4Bharat)**
- Language pair: Hindi (Devanagari) $\rightarrow$ English (Roman)
- Training data: $\leq$ 80k examples (uniform random subsampling)
- Validation set was the remainder from the train set in our case, and test set from Aksharantar
- Potential cleaning steps:
  - Remove duplicates (We did not check for duplicates because our first-try results were already satisfactory)
  - Normalized Unicode characters (remove accents/diacritics)
  - Tokenized at character level

Non-ML Baseline (hardcoded rule-based mapping)

```
CHAR_MAP = {
    # Vowels
    'अ': 'a', 'आ': 'aa', 'इ': 'i', 'ई': 'ii', 'उ': 'u', 'ऊ': 'uu',
    'ऋ': 'ri', 'ॠ': 'rii', 'ऌ': 'lri', 'ए': 'e', 'ऐ': 'ai',
    'ओ': 'o', 'औ': 'au',

    # Consonants
    'क': 'ka', 'ख': 'kha', 'ग': 'ga', 'घ': 'gha', 'ङ': 'nga',
    'च': 'cha', 'छ': 'chha', 'ज': 'ja', 'झ': 'jha', 'ञ': 'nya',
    'ट': 'ta', 'ठ': 'tha', 'ड': 'da', 'ढ': 'dha', 'ण': 'na',
    'त': 'ta', 'थ': 'tha', 'द': 'da', 'ध': 'dha', 'न': 'na',
    'प': 'pa', 'फ': 'pha', 'ब': 'ba', 'भ': 'bha', 'म': 'ma',
    'य': 'ya', 'र': 'ra', 'ल': 'la', 'ळ': 'la', 'व': 'va',
    'श': 'sha', 'ष': 'sha', 'स': 'sa', 'ह': 'ha',
```

Table: Baseline Transliteration Metrics

Metric	**Value**
Top-1 Accuracy (ACC)	0.0014 (0.14%)
Mean Precision	0.6379
Mean Recall	0.9419
Mean F1 (Fuzziness)	0.7561

LSTM Based Transliteration

- Encoder–decoder model without attention, single layer LSTM
- Embedding dimension of 128, Hidden dimension of 256
- Trained for 5 epochs on 80k input sequences (until validation and training loss were just about to diverge). Used a batch size of 64 and Adam optimizer with lr=1e-3
- Trained using teacher forcing (50%)
- Compared:
  - Greedy decoding
  - Beam search (beam width = 3)
- Referred to SaLP (Jurafsky/Martin) for LSTM theory as a refresher from A1
- **Demo**

LSTM Based Transliteration

Metric	**LSTM (Greedy)**	**LSTM (Beam)**
Top-1 Accuracy (ACC)	0.3592	0.3683
Mean Precision	0.9115	0.9143
Mean Recall	0.8927	0.8942
Mean F1 (Fuzziness)	0.8983	0.9005

Table: Comparison of LSTM transliteration performance using Greedy vs Beam Search decoding.

LSTM Based Transliteration (Failures)

```
Words for which F1-score < 0.5:
Hindi: एनसीआरडब्ल्यूसी      Gold: enseeaardablyusee    Pred: ncrbducy
Hindi: डब्ल्यूसीसीएल        Gold: wccl                 Pred: dbulciale
Hindi: एचजेडटीसी            Gold: echajedateesee       Pred: hjdtc
Hindi: ईडब्लूएस             Gold: ews                  Pred: eadbluse
Hindi: आईआरडब्ल्यू          Gold: irw                  Pred: irdblue
Hindi: यूजी                  Gold: ug                   Pred: uji
Hindi: एआईआरटीयू            Gold: eaaiaartiyoo         Pred: airtu
Hindi: लफ्ज                  Gold: laphj                Pred: lufz
Hindi: आईआरपीटीसी           Gold: aaeeaarpitisi        Pred: irpttc
Hindi: बीआईआईटीएम           Gold: beeaaeeaaeeteeem     Pred: biitm
Hindi: डीडब्ल्यूएफ          Gold: dwf                  Pred: ddublf
Hindi: क्यूसैक्स            Gold: kyusaiks             Pred: cusex
Hindi: एनडब्ल्यूसी          Gold: nwc                  Pred: andblucy
Hindi: क़ि                   Gold: ki                   Pred: qui
Hindi: सीडब्ल्यूएस          Gold: cws                  Pred: seadbluse
Hindi: डब्ल्यूएचएल          Gold: dablyooechel         Pred: dbul
Hindi: एनसीआरडब्ल्यूसी     Gold: enseeaardablyoosee   Pred: ncrbducy
Hindi: क्यूडब्ल्यूवीजीए     Gold: qwvga                Pred: qudblogy
Hindi: एमपीडब्ल्यूपीपीसीएल Gold: empeedablyoopeepeeseeel Pred: mpddblpc
Hindi: पीडब्ल्यूआई          Gold: pwi                  Pred: pdebuia
Hindi: पीएचडब्ल्यूसीएस      Gold: peeechdablyuseees    Pred: phdbsc
Hindi: डीओडब्ल्यू           Gold: dow                  Pred: dodblue
Hindi: ईआरजी                Gold: eeaarjee             Pred: erg
Hindi: क्यूडब्ल्यूवीजीए     Gold: kyoodablyooveejeea   Pred: qudblogy
Hindi: पीडब्ल्यूटी          Gold: pwt                  Pred: peedbluty
Hindi: हुक़ूक़               Gold: hukhookh             Pred: huquq
```

Transformer Based Transliteration

- Implemented a 2-layer transformer encoder-decoder
- Used local attention (window masking)
- Trained for 10 epochs, Adam optimizer (lr=3e-4)
- Used two decoding strategies (greedy and beam)
- Referred to SaLP (Jurafsky/Martin) to understand the architecture
- Demo

Transformer Based Transliteration

Metric	**Greedy**	**Beam**
Top-1 Accuracy (ACC)	0.3565	0.3586
Mean Precision	0.9379	0.9352
Mean Recall	0.9048	0.9073
Mean F1 (Fuzziness)	0.9170	0.9169

Table: Comparison of Transformer transliteration performance using Greedy vs Beam Search decoding.

Transformer Based Transliteration (Failures)

Source	Reference	Greedy Pred	Beam Pred	G-F1	B-F1
एनडब्ल्यूसी	endablyoosee	nwc	nwc	0.2667	0.2667
एनडब्ल्यूसी	endablyusee	nwc	nwc	0.2857	0.2857
डब्ल्यूईएफ	dablyooeeef	wif	wif	0.2857	0.2857
एनसीआरडब्ल्यूसी	enseeaardablyoosee	ncrwc	ncrwc	0.3043	0.3043
डब्ल्यूईएफ	dablyueeef	wif	wif	0.3077	0.3077
डब्ल्यूपीडी	dablyoopeedee	wpd	wpd	0.3125	0.3125
पीडब्ल्यूटी	peedablyootee	pwt	pwt	0.3125	0.3125
डीडब्ल्यूटी	deedablyootee	dwt	dwt	0.3125	0.3125
आईआरडब्ल्यू	aaiaardablyoo	irw	irw	0.3125	0.3125
एनसीआरडब्ल्यूसी	enseeaardablyusee	ncrwc	ncrwc	0.3182	0.3182
डब्ल्यूएचएल	dablyooechel	whl	whl	0.3333	0.3333
आईआरडब्ल्यू	aaiaardablyu	irw	irw	0.3333	0.3333
डीडब्ल्यूटी	deedablyutee	dwt	dwt	0.3333	0.3333
डीडब्ल्यूएन	deedablyooen	dwn	dwn	0.3333	0.3333
पीएचडब्ल्यूसीएस	peeechdablyooseees	phwcs	phwcs	0.3478	0.3478
पीडब्ल्यूआईडी	peedablyooaaidee	pwid	pwid	0.3500	0.3500
बीआईआईटीएम	beeaaeeaaeeteeem	bitm	bitm	0.3500	0.3500
डीडब्ल्यूएन	deedablyuen	dwn	dwn	0.3571	0.3571
डब्ल्यूपीडी	dablyoopidi	wpd	wpd	0.3571	0.3571
डब्ल्यूपीडी	dablyupeedi	wpd	wpd	0.3571	0.3571

LLM Based Transliteration

- Used GPT-5 for transliteration
- Tried various temperature (T) and top-p settings
- Prompt:
  "You are a transliteration assistant."
  "Transliterate the following Hindi (Devanagari) text into English (Roman script), preserving pronunciation."
  "Do not use accents or diacritics."
  "Return only the transliteration."
  "Hindi: {hindi text} English:"
- We batched inputs in groups of 50 to save tokens by avoiding wasteful reprompting
- **Demo**

LLM Based Transliteration

Table: LLM-based Transliteration Metrics for different top_p values

Metric	**p = 0.1**	**p = 0.5**	**p = 1.0**
Top-1 Accuracy (ACC)	0.2965	0.2795	0.2745
Mean Precision	0.9222	0.9056	0.9180
Mean Recall	0.8637	0.8474	0.8596
Mean F1 (Fuzziness)	0.8873	0.8701	0.8834

Table: LLM-based Transliteration Metrics for different temperature (T) values

Metric	**T = 0.1**	**T = 0.5**	**T = 1.0**
Top-1 Accuracy (ACC)	0.2948	0.2855	0.2762
Mean Precision	0.9178	0.9157	0.9214
Mean Recall	0.8587	0.8581	0.8649
Mean F1 (Fuzziness)	0.8824	0.8810	0.8879

LLM Based Transliteration (Failures)

```
Words for which F1-score < 0.5:
Hindi: एनसीआरडब्ल्यूसी      Gold: enseeaardablyusee    Pred: NCRWCS
Hindi: डीडब्ल्यूएन          Gold: deedablyuen          Pred: DWN
Hindi: एचजेडटीसी             Gold: echajedateesee       Pred: HZTC
Hindi: पीडब्ल्यूआईडी        Gold: peedablyooaaidee     Pred: PIDW
Hindi: यूजी                    Gold: ug                   Pred: yuji
Hindi: डब्ल्यूआईएन           Gold: win                  Pred: dablyoo aai en
Hindi: एआईआरटीयू             Gold: eaaiaartiyoo         Pred: AIRTU
Hindi: एलऐंडटी               Gold: elendtee             Pred: L&T
Hindi: आईआरपीटीसी            Gold: aaeeaarpitisi        Pred: IRPTC
Hindi: बीआईआईटीएम           Gold: beeaaeeaaeeteeem     Pred: BIITM
Hindi: एनडब्ल्यूसी          Gold: endablyusee          Pred: NWC
Hindi: ऑक्सीज                 Gold: oxys                 Pred: oksij
Hindi: सीडब्ल्यूएस           Gold: cws                  Pred: si-dabluyes
Hindi: डब्ल्यूपीडी          Gold: dablyoopeedee        Pred: WPID
Hindi: आईआरपीटीसी             Gold: aaiaarpitisi         Pred: IRPTC
Hindi: एनआरडीडब्ल्यूपी      Gold: nrdwp                Pred: enaradyudablyupi
Hindi: क्यूडब्ल्यूवीजीए     Gold: qwvga                Pred: kyudablyuvijiye
Hindi: डब्ल्यूईएफ             Gold: dablyueeef           Pred: WEF
Hindi: एमपीडब्ल्यूपीपीसीएल  Gold: empeedablyoopeepeeseeel Pred: MPDWPPCL
Hindi: पीएचडब्ल्यूसीएस      Gold: peeechdablyuseees    Pred: PHWCS
Hindi: डब्ल्यूपीडी          Gold: dablyoopidi          Pred: wpd
Hindi: बीडब्ल्यूटीएस        Gold: beedablyooteees      Pred: BWTS
Hindi: ईआरजी                  Gold: eeaarjee             Pred: ERG
Hindi: क्यूडब्ल्यूवीजीए     Gold: kyoodablyooveejeea   Pred: QWVGA
Hindi: आईआरडब्ल्यू            Gold: aaiaardablyu         Pred: IRW
Hindi: एनएसआईयू               Gold: enesaaeeyoo          Pred: NSIU
```

Comparison of Models

We find that well-trained LSTM and transformer models outperform LLMs in both top-1 accuracy and soft F1 because of task-specialization. However, we would like to challenge this conclusion to an extent because this Aksharantar dataset contains many incorrect gold labels as well. A solution to this would be to have overlapping random subsets of human-evaluations to weed out incorrect transliterations with a confidence score. This would of course require some manual labor to achieve a higher quality dataset.

Analysis and Observations

- Common error patterns:
  - Acronyms
  - Nasal sounds often missed
  - Ambiguous vowels (lahar vs leher)
- Local attention improved shorter word accuracy
- Beam search slightly outperformed greedy decoding
- LLMs much slower (overkill for this task)

References

- Aksharantar Dataset: `https://huggingface.co/datasets/ai4bharat/Aksharantar`