Phoneme Bridges: Leveraging Phonetic Similarity for Low-Resource Language Understanding

Chris Ge, Brian Le, Daria Kryvosheieva

Cross-lingual Transfer

- **90%** of languages, with over **1 billion** speakers, are currently **ignored** by NLP technologies, due to severe disparities in the availability of training data.
- Some languages have **different writing systems** but **similar words and pronunciation**

Example: Different Script, Same Pronunciation

Hindi: **क्या आप ठीक हैं?** (Kyaa aap theek hain?)
Urdu: **کیا آپ ٹھیک ہیں؟** (Kyaa aap theek hain?)

"Are you okay?"

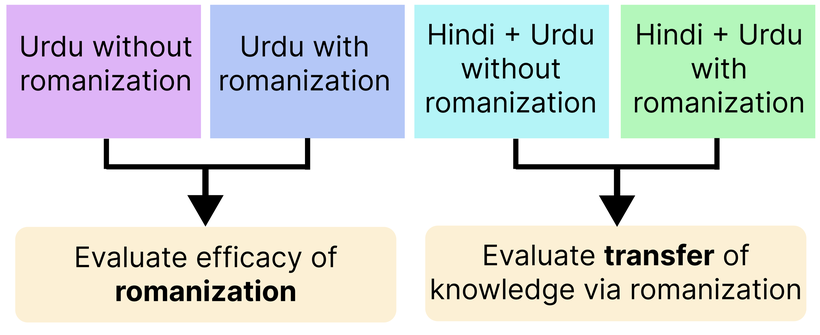
Can we use the similarity in pronunciation to transfer LLM learning from a higher-resource language (Hindi) to a lower-resource language (Urdu)?

Idea: Capture pronunciation with **romanizations**

- Fine-tune for task on romanized Hindi first, then fine-tune on romanized Urdu

Experiments

- Fine-tune **four** models and make **two** comparisons



- Across **two** tasks (from the Google Xtreme dataset)

Part-of-Speech Tagging | Named-Entity Recognition

Training Pipeline

Models without romanization/Hindi omit the romanization/Hindi steps

A: Obtain Romanizations — ai4bharat transliteration

Hindi:	मुझे	पानी	चाहिए
romanizations:	mujhe	paani	chahiye
Urdu:	چاہیے	پانی	مجھے
romanizations:	chahiye	paani	mujhe

(Urdu is written right to left)

↓ Data

B: Hindi Intermediate Fine-Tuning

मुझे paani चाहिए

↓ +task labels

- 25% random **replacement** by romanization

mBERT

- **Goal**: learn useful romanization representation for task

↓ Intermediate mBERT

C: Urdu Fine-Tuning

مجھے پانی چاہیے mujhe paani chahiye

↓ +task labels

- Full **concatenation** with romanization

Intermediate mBERT

- **Goal**: learn task in Urdu

↓

final mBERT

- **Evaluate** on Urdu task

Results

Macro-F1 scores

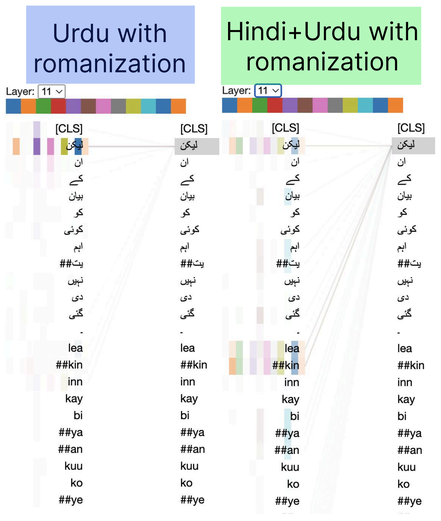
Model	POS Tagging Score	NER Score
$\text{Urdu}_{\text{plain}}$	0.8700	0.9770
$\text{Urdu}_{\text{romanization}}$	**0.8728**	**0.9780**
$\text{Hindi+Urdu}_{\text{plain}}$	0.8702	0.9763
$\text{Hindi+Urdu}_{\text{romanization}}$	**0.8735**	**0.9788**

Statistical Testing

Comparison	Task	Mean Difference	95% Confidence Interval	P-value
$\text{Urdu}_{\text{romanization - plain}}$	POS	0.0029	[-0.0016, 0.0071]	0.2180
	NER	0.0010	[-0.0066, 0.0079]	0.7680
$\text{Hindi+Urdu}_{\text{romanization - plain}}$	POS	0.0035	[-0.0019, 0.0092]	0.2200
	NER	0.0026	[-0.0038, 0.0095]	0.4100

Attention visualization (using BertViz)

Urdu with romanization | Hindi+Urdu with romanization



Attention paid to the correct romanizations only on the right side model for UD-POS dataset

→ Hindi fine-tune helps learn Urdu romanizations, but not task

For PAN-X dataset, models had similar attention

→ Sometimes Urdu fine-tune is sufficient to learn romanizations

Conclusions

- On all NLP tasks we investigated, our approach led to **improvement**, but the improvement was **not statistically significant**.
- Our inspection of mBERT attentions with BertViz suggests the model **correctly learned the associations** between words in the original script and their romanizations, but **failed to exploit** the romanizations for the tasks.
- Future work could use our approach for applications that require learning romanized representations of words in non-Latin scripts.

https://github.com/brian224code/phoneme-bridges