Limited Attention: Investigating Transformer Models' Zero-Shot Cross-Lingual Transfer Learning with Urdu Named Entity Recognition

Anwesha Mukherjee

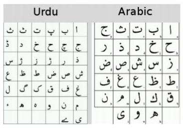
Overview

Low Resource Language Challenge:
Many languages lack tagged or processable data to train or even fine-tune models for specific tasks.

Linguistic Transfer Question:
Can we fine-tune a model for a task in a similar language to work for a low resource language?
Idea:
- Fine-tune multilingual models using few-shot learning:
  - Training data from a high-resource related language
  - Validation and testing data in the target language
- Compare performance to models finetuned with target language training data.

Why Urdu to Model the Problem?
- Morphological richness with ambiguous language composition
- No capitalization
- Script (typological) vs. Vocabulary (morphological) Question
  - Indic language and massive shared vocabulary with Hindi
  - Arabic/Farsi derived Script

Hindi Word	After Transliteration	Conventional way
ब्लैकमेल	بلیکمیل	بلیک میل
बिजलीघर	بجلیگھر	بجلی گھر
बातचीत	باتچیت	بات چیت



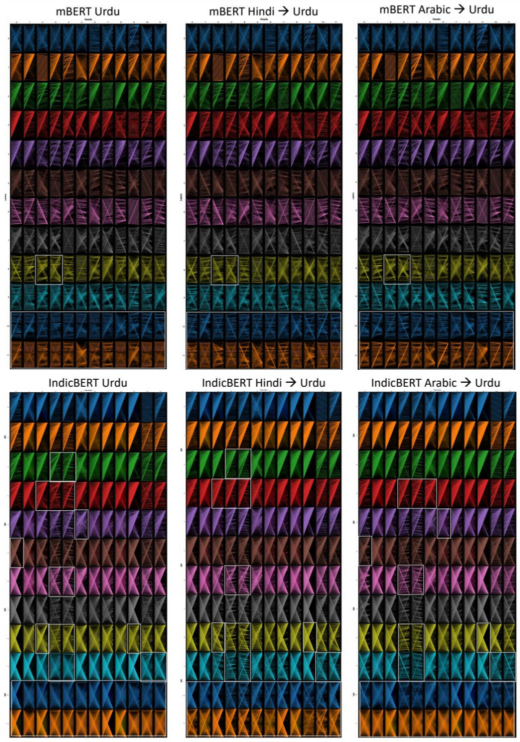
Background

mBERT: Multilingual BERT
- SoTA language model pretrained on monolingual corpora of 104 languages
- Suitable for typological transfer and morphological transfer

IndicBERT: Indic Language multilingual ALBERT
- 12 languages (11 Indic and Indian English)
- Modified hyperparameters with smaller model

Data: WikiANN Named Entity Recognition Urdu Data
- used in IndicGLUE evaluation of both models
- 3 tags:
  - Person
  - Organization
  - Location

Attention Analysis for Transformer Models

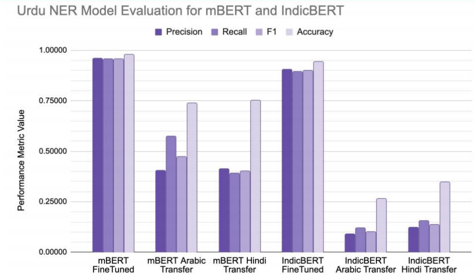


- mBERT has better distributive attention
- Final 2 layers are instrumental to performance dropoff
- Intermediate attention loses spread with transfer learning

Notable References

- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jesse Vig. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Comparative Model Performance



- IndicBERT – no Arabic Alphabet
- Morphological Similarities – Higher Accuracy
- Typological Similarities – Higher f1
- Direct Fine-Tuning Converges within 7 epochs
- Transfer Fine-Tuning Does not Approach

Potential Causes: Architectural Differences

Difference	mBERT	IndicBERT	
Dropout	0.1	0	0 dropout caters to sequence classification and **overfits to training language**
Model Size	BERT	ALBERT	ALBERT has 9x fewer parameters and **6x fewer embedding layers**
Embeddings	104 langs	12 langs	IndicBERT has **no unit embeddings** for the **Arabic** Alphabet
Tokenizer	WordPiece	SentencePiece	**WordPiece:** maximize the likelihood of the training data. **SentencePiece:** pair frequency. Wordpiece had higher rate of merges.

Insights on Cross-Lingual Transfer NER

Token Classification – relies on mix of context and character embeddings

Attention – final layers dictate sequential units

Typological similarities dominate efficacy

Blocks Frozen – Model performs better finetuning all layers rather than just the classifier.

Early Stopping – In transfer learning, the gradient fluctuates largely so early stopping ends training prior to the model's optimal performance.

Next Steps

1. Pretrained Roberta – UrBERTo from scratch (without compute limitations)
2. Mixed few-shot transfer learning (typological + morphological)