**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?

**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

<table>
<thead>
<tr>
<th>Urdu</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>कृति</td>
<td>میں</td>
</tr>
<tr>
<td>फातिमा</td>
<td>خوشحال</td>
</tr>
<tr>
<td>साहिल</td>
<td>دیکت</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>PERSON</th>
<th>ORG</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ryan</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ryan attended the Russian Academy of Sciences in India.

---

**Attention Analysis for Transformer Models**

- mBERT Urdu
- mBERT Hindi > Urdu
- mBERT Arabic > Urdu

**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**

- Dropout
  - mBERT
  - IndicBERT
  - 0 dropped tokens 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
  - 32 langs
  - IndicBERT has no unit embeddings for the Arabic Alphabet

- Tokenizer
  - WordPiece
  - SentencePiece
  - SentencePiece: maximizes the likelihood of the training data
  - WordPiece: pair frequent words and had higher rate of merges

**Insights on Cross-Lingual Transfer NER**

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

**Next Steps**

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