# Final Project Evaluation

Sneha Oram, 23M2159, 1, CMInDS
A Snegha, 23M2160, 1, CMInDS
K Hemanth, 23M2164, 1, CMInDS

# Problem Statement

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451-462.
- Input - Monolingual word embedding of source and target language.
- Output - Bilingual word embedding.

# Progress

- We trained Punjabi, Tamil, Telugu monolingual word embedding model using **AI4Bharat-IndicNLP Dataset** for **Punjabi**, **Tamil**.
- The monolingual word embedding training is done using **200MB, 500MB and 1GB** of the original Punjabi (8GB) and Tamil (11GB) datasets.
- Now using the monolingual word embeddings we trained a bilingual word embedding for Punjabi - Tamil on 4 set of train dictionaries - 25, 75, 100, 1000 word pair.

# Progress

- Collected 1100 word-pair dictionary for Punjab - Tamil.
- Evaluated on bilingual lexical induction - Punjabi to Tamil
- Top 5 predictions are given by model.
- Test set size = 100 word pair

# Bilingual Lexical Induction - Accuracy

| Dataset - > | 200 MB | 500 MB | 1 GB |
|---|---|---|---|
| 1000 Word pair | 13.68% | 11.58% | 13.68% |
| 100 Word pair | 3.16% | 3.16% | 10.53% |
| 75 Word pair | 2.11% | 2.11% | 7.37% |
| 25 Word pair | 0% | 0% | 0% |

Test size = 100 word pair.

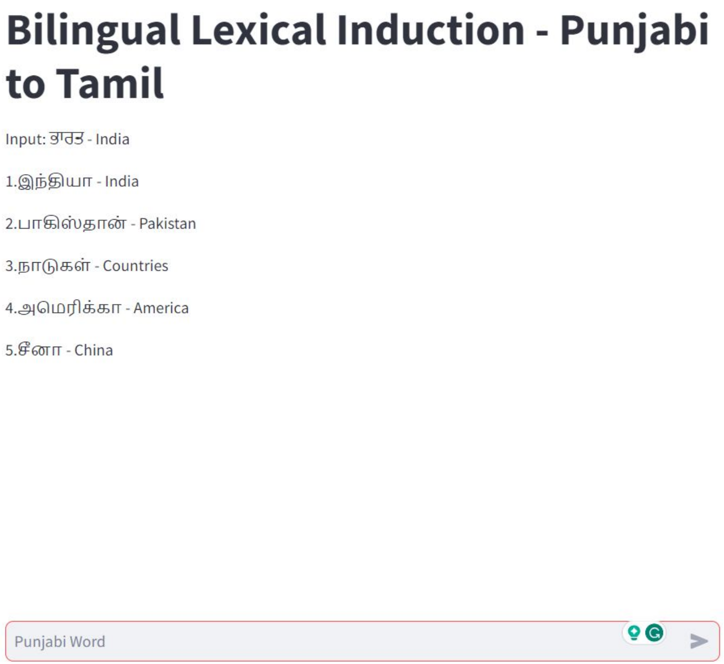
# Analysis

- Paper hypothesis: Bilingual lexical induction performance with 25 training word pair is comparable to 5000 training word pair.
- Observation: Punjabi - Tamil bilingual word embedding need larger monolingual dataset and more bilingual word pairs.
- Model is performing better with 1 GB dataset, however it is taking very long time in execution compared to other datasets.

# Analysis

- As the families are different, we tried increasing model performance with increasing monolingual dataset size, and bilingual word pair sets
- Model is predicting in context.
- Model didn’t see few morphological variation of words.
- Two different words with same meaning are correctly identified and same result is generated.
  - Ex - ('ஆறு', 'ஆற்றில்') ('நதி', 'ஆற்றில்')

# Demo



## Bilingual Lexical Induction - Punjabi to Tamil

Input: ਭਾਰਤ - India

1.இந்தியா - India

2.பாகிஸ்தான் - Pakistan

3.நாடுகள் - Countries

4.அமெரிக்கா - America

5.சீனா - China

# Thank You