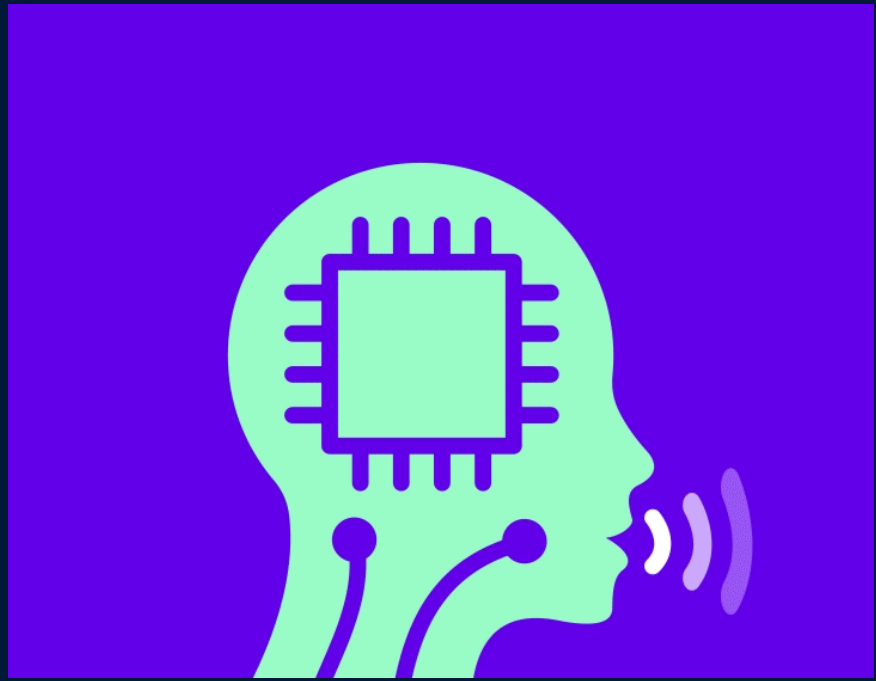
Cross-Lingual Word Embeddings: A Comparative Study of Bangla and English Word Embeddings Trained on Monolingual Corpora



CSE499A: **SENIOR DESIGN I**

BY

Mahobub Shahoriar Siam 2013302642
Tashnia Tabassum 2011496642

The problem statement

This project aims to delve into the concept of extracting semantic similarities from Bangla and English word vectors, derived from distinct monolingual corpora. The objective is to facilitate knowledge transfer between these languages, investigating the notion that various languages share identical meanings. The primary goal is to substantiate the proposition that semantic significance remains consistent across these diverse linguistic contexts. By demonstrating the alignment of semantic meanings, this study endeavors to establish a deeper understanding of the universality of meaning across different languages.

Related works

→ Hong Jin Kang et al. compared popular word embeddings for monolingual English word sense disambiguation (WSD) and cross-lingual WSD in Chinese, achieving state-of-the-art results without retraining and revealing limitations of basic LSTM networks for the tasks.
Ref:https://aclanthology.org/W16-4905/?fbclid=IwAR0eK3MudhEWo5ypLhyMg5aPqi1-uxcquPIcCfQ-YKLsxFUggoOSuZ5R4No

→ Jiapeng et al. conducted rigorous textual analysis and proposed a comprehensive classification system which had crucial found out two key aspects of it which were: text distance and representation. For text distance, they evaluated many distance metrics such as cosine distance, euclidian distance, Hamming, Wesserstein distances etc.
Ref:https://www.mdpi.com/2078-2489/11/9/421?fbclid=IwAR1E_BhM1RnCS19wELrgIGPnpXQ8pu_x4lU7N9tQgYWzRDKriz-YZZHPiQU

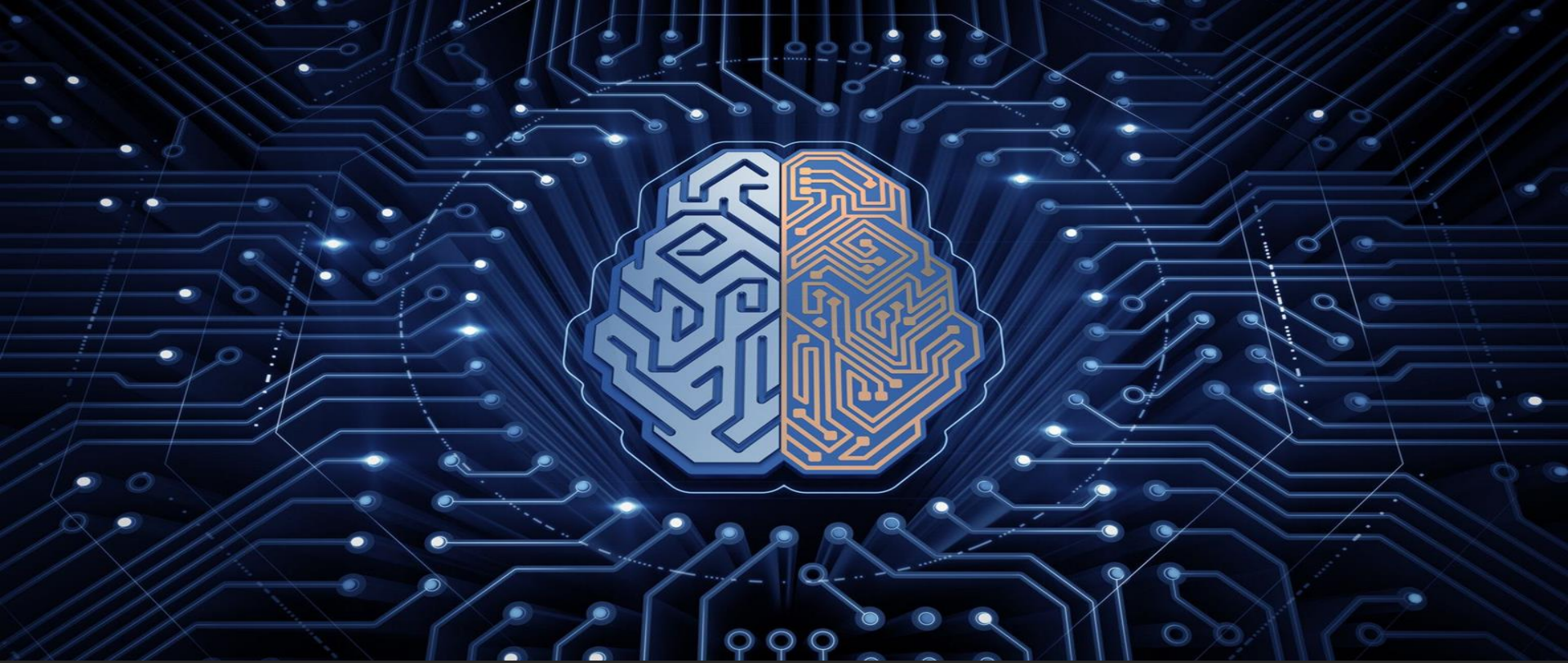
A working plan

This project aims is a research study that will compare the similarity distances between translations of words. We will focus on utilizing the largest Bangla newspaper dataset for training word embeddings using word2vec and conducting a comparative analysis with the English counterpart already trained on the Google News dataset using word2vec. By undertaking this comparative study, we will try to understand the applicability of cross-lingual word embeddings, especially with Bangla being a low resource language. Then, we will use the cosine similarity to evaluate and cross-check between selected words in Bangla and English. Later on, we plan to use more distance metrics for further evaluation and more different state of the art approaches.

Dataset Statistics

→ Largest Newspaper Dataset of Bangladeshi Bangla Newspapers, containing articles from Ittefaq, Jugantor & Prothom Alo. **Sourced from: https://www.kaggle.com/datasets/ebiswas/bangla-largest-newspaper-dataset**

Dataset Total Size	22GB (Used: 12GB due to lack of computational resources)
Total Records	2.2 Million Articles
Total Words from the partial dataset used	291446363
File Type	JSON



Thank You!