ross Lingual Embeddings for English to Hindi

Mukundan

April 7 2025

bstract

This research explores cross-lingual word embedding alignment between English and Hindi languages using FastText pre-trained embeddings. The study aims to bridge the semantic gap between these typologically distinct languages by creating a robust mapping between their respective embedding spaces.

We implement and evaluate two primary approaches for cross-lingual alignment: supervised Procrustes alignment and unsupervised adversarial training. The study utilizes the MUSE bilingual lexicon for training and evaluation, with vocabulary limited to 100,000 most frequent words in each language to ensure computational efficiency while maintaining coverage of commonly used terms.

Our methodology incorporates several advanced techniques including Procrustes analysis for orthogonal mapping, Cross-domain Similarity Local Scaling (CSLS) for improved nearest neighbor retrieval, and adversarial training with a discriminator network to refine the alignment quality. The implementation was carried out on Amazon SageMaker using an ML.G5.2xlarge GPU instance with 24GB VRAM and 500GB storage capacity, enabling efficient processing of large embedding matrices and rapid training of the neural network components.

The experimental results demonstrate the effectiveness of both approaches, with precision@1 reaching 28.5

This research contributes to the field of cross-lingual embedding alignment by providing detailed insights into the challenges and solutions for aligning embedding spaces between English and Hindi. The established framework and methodology can serve as a foundation for future work in cross-lingual natural language processing tasks involving these languages.

Overview

Repository location The dataset and implementation code are openly available on GitHub at

https://github com/lucifer1702/Embeddings_English_Hindi

ontext This research was conducted as part of an assignment for Sarvam AI focusing on cross-lingual natural language processing, specifically addressing the challenge of aligning word embeddings between English and Hindi languages. The implementation utilizes several key methodologies and datasets:

The implementation was carried out using Amazon SageMaker's ML.G5.2xlarge GPU instance with 24GB VRAM and 500GB storage capacity, enabling efficient processing of large embedding matrices and neural network training. The research builds upon several seminal works in cross-lingual embedding alignment:

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135-146.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H. (2017). Word Translation Without Parallel Data. arXiv preprint arXiv:1710.04087.

Artetxe, M., Labaka, G., Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.