ross Lingual Embeddings for English to Hindi

Mukundan

April 8 2025

Abstract

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 e ectiveness of both approaches, with precision@1 reaching 28.5 percent and precision@5 achieving 53.1 percent on the test set. An ablation study examining the impact of training dictionary size, ranging from 2,000 to 8,000 word pairs, reveals consistent improvement in alignment quality with increased training data, with precision@1 improving from 16 percent to 27 percent.

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.

1 Overview

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

https://github com/lucifer1702/Embeddings_English_Hindi

Context 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.

2 Methodology

2 1 Data Preparation

2.1.1 Pre-trained Word Embeddings

This study utilizes pre-trained FastText embeddings for both English and Hindi due to their high quality and e ectiveness for morphologically rich languages. The 300-dimensional vector models (cc.en.300.bin and cc.hi.300.bin) are loaded directly from the FastText repository, providing comprehensive representations trained on Common Crawl and Wikipedia data.

2.1.2 Vocabulary Limitation

To ensure computational efficiency while maintaining robust representation, we limit the vocabulary to the 100,000 most frequent words in each language. This filtering process leverages FastText's frequency-ordered word lists, providing coverage of common terminology while significantly reducing memory requirements. This limitation is critical for efficient processing on the AWS SageMaker ML.G5.2xlarge GPU instance used for implementation.

2.1.3 Bilingual Lexicon

Word translation pairs are sourced from the MUSE dataset, which serves as the gold standard for cross-lingual evaluation. The dataset is split into:

Training set (en-hi.0-5000.txt): Used for learning the mapping between embedding spaces

Testing set (en-hi.5000-6500.txt): Used for evaluation of translation accuracy

Only word pairs where both words exist in our limited vocabulary are retained, resulting in 8,130 training pairs and 1,600 testing pairs. This filtering ensures alignment is based on words with reliable embeddings in both languages.

2 2 Embedding Alignment

2.2.1 Supervised Procrustes Alignment

The core alignment approach implements Procrustes analysis, which finds an optimal orthogonal transformation between embedding spaces. This method:

Preserves the geometric structure of the original embeddings

Maintains angles and distances between word vectors

Creates a linear mapping that transforms English embeddings to match Hindi embeddings

The process involves:

Extracting source (English) and target (Hindi) word vectors for all word pairs in the training lexicon

Normalizing all vectors to unit length using L2 normalization

Applying singular value decomposition to find the optimal orthogonal transformation matrix

2.2.2 Cross-Domain Similarity Local Scaling CSLS)

To address the hubness problem (where some target words become nearest neighbors of many source words), we implement CSLS. This technique:

Adjusts similarity scores based on the average similarity of neighbors in both source and target spaces

Reduces the bias toward hubs in the target space

Improves the precision of nearest neighbor retrieval

2 3 Evaluation Framework

2.3.1 Word Translation Task

The aligned embeddings are evaluated on word translation retrieval:

For each English source word, find nearest neighbors in the Hindi embedding space

Compare the retrieved Hindi words with the gold standard translations

2.3.2 Precision Metrics

Translation accuracy is measured using:
Precision@ : Percentage of correct translations found as the top result
Precision@5: Percentage of correct translations found within the top 5 results

2.3.3 Cosine Similarity Analysis

To understand the semantic relationships between translation pairs:
Calculate cosine similarities between true translation pairs
Examine highest and lowest similarity pairs to identify patterns and challenges

2.3.4 Ablation Study

To determine the impact of training data size on alignment quality:

Create subsets of the bilingual lexicon: 2,000, 4,000, 6,000, and 8,000 word pairs

Train separate alignment models for each subset

Compare precision metrics across models to quantify the relationship between lexicon size and translation accuracy

This comprehensive methodology enables a thorough investigation of cross-lingual word embedding alignment between English and Hindi, addressing both supervised and unsupervised approaches while providing detailed analysis of factors a ecting alignment quality.

2 4 Unsupervised Alignment Method

The unsupervised alignment method implements a novel approach that eliminates the need for bilingual dictionaries through an adversarial framework combined with strategic refinement techniques. At its core, the method employs a discriminator network that attempts to distinguish between mapped source embeddings and actual target embeddings, while simultaneously training a mapping function that aims to deceive this discriminator by generating increasingly convincing translations. This adversarial process is constrained by orthogonality requirements that preserve the structural integrity of the embedding spaces, ensuring that semantic relationships remain intact during transformation. Following initial convergence, the mapping undergoes systematic refinement through Cross-domain Similarity Local Scaling (CSLS), which mitigates the hubness problem inherent in high-dimensional spaces, followed by the extraction of a synthetic dictionary based on mutual nearest neighbors between the language spaces. This synthetic dictionary then enables a Procrustes refinement step, leveraging the strongest identified translation pairs to further optimize the mapping. The entire process operates in a fully unsupervised manner, making it particularly valuable for language pairs like English-Hindi where extensive parallel resources may be limited.

3 Results and Inference

3 1 Translation Performance

The evaluation of our cross-lingual embedding alignment system reveals promising results in enabling word translation between English and Hindi. The supervised Procrustes alignment method demonstrates considerable e ectiveness with a Precision@1 score of 28.5 percent and a Precision@5 score of 53.1 percent when evaluated on the MUSE test dictionary comprising 1,600 word pairs. These metrics indicate that for more than one-quarter of English query words, the correct Hindi translation appears as the top result, while for over half of the queries, the correct translation is found within the top five candidates.

Sample translations showcase the system's capabilities across di erent semantic domains. When translating the word "hello" to Hindi, the system correctly identifies "" (halo) and "" (hello) among its top candidates, along with semantically related greetings like "" (namaste). Similarly, for the English word "name," the system produces "" (naam) as the top result, accurately capturing the semantic equivalence despite morphological di erences between the languages.

3 2 Semantic Similarity Analysis

Analysis of cross-lingual semantic similarity through cosine distances yields interesting insights into the alignment quality. The five most similar word pairs identified by our system demonstrate strong alignment between concepts that have either direct transliterations or clear semantic equivalence:

English Word	Similarity
lecture	0.6854
kilometers	0.6737
consultant	0.6331
aggressive	0.6236
extraordinary	0.6007

Table 1: Top 5 word pairs with highest cosine similarity after alignment

These results indicate that the alignment process is particularly e ective for technical terms, measurement units, and descriptive adjectives. The high similarity score for "kilometers-" (0.6737) reflects the phonetic similarity and conceptual equivalence, while more abstract concepts like "aggressive-" (0.6236) demonstrate the system's ability to capture semantic relationships beyond surface-level similarities.

3 3 Ablation Study: Impact of Training Dictionary Size

The ablation study examining the relationship between bilingual lexicon size and alignment quality reveals a clear correlation between training data volume and translation accuracy. Results demonstrate consistent improvements in both Precision@1 and Precision@5 metrics as the training dictionary size increases:

Dictionary Size	Precision@1	Precision@5
2 000	0.16	0.37
4 000	0.23	0.46
6 000	0.25	0.49
8 000	0.27	0.51

Table 2: Translation accuracy metrics across varying training dictionary sizes

The most significant improvement occurs when expanding from 2,000 to 4,000 word pairs, where Precision@1 increases by 7 percentage points (16 percent to 23 percent) and Precision@5 improves by 9 percentage points (37 to 46). The gains become more incremental with further increases in training data, suggesting a logarithmic relationship between dictionary size and alignment quality. This pattern indicates that while more training data is beneficial, there are diminishing returns after reaching approximately 6,000 word pairs for this language pair.

3 4 Comparative Analysis: Supervised vs Unsupervised Methods

The comparison between supervised Procrustes alignment and unsupervised adversarial training with CSLS provides valuable insights into the trade-o s between these approaches. Both methods achieved comparable performance on sample translations, although the supervised approach maintained a slight edge in overall precision.

The results from the unsupervised implementation show the robustness of the adversarial approach, which achieved competitive performance without requiring pre-existing bilingual dictionaries. This achievement is particularly notable given the typological di erences between English and Hindi, demonstrating the method's adaptability across language families.

During the unsupervised training process, we observed steady improvement in both discriminator and generator losses across epochs. The discriminator loss decreased from 1.3725 to 1.3502 between the first and second adversarial epochs, while the generator loss remained relatively stable (0.7306 to 0.7415), indicating e ective adversarial learning dynamics.

The CSLS refinement process successfully addressed the hubness problem, improving nearest neighbor retrieval by adjusting similarity scores based on the neighborhood characteristics of both source and target embedding spaces. This refinement was particularly crucial for Hindi, where certain words might otherwise dominate as translation candidates due to their central position in the embedding space.

Overall, these results demonstrate the viability of both supervised and unsupervised approaches for cross-lingual word embedding alignment between English and Hindi, with the choice between methods depending on the availability of bilingual resources and specific application requirements.

4 Future Improvements

Several avenues exist for enhancing the cross-lingual embedding alignment system presented in this study. A promising direction would be to incorporate contextual embeddings from transformer-based models like BERT or XLM-R, which could capture polysemy and context-dependent meanings more e ectively than the static FastText embeddings currently employed. The alignment process could be further refined by implementing iterative refinement techniques that progressively improve the mapping quality through multiple cycles of dictionary induction and realignment. Additionally, exploring multi-step mapping approaches that utilize a high-resource pivot language could potentially improve results for the English-Hindi pair by leveraging intermediate representations. Performance could also be enhanced by expanding the training dictionary with domain-specific terminology or by employing subword-level alignment to better handle the morphological complexity of Hindi. From a methodological perspective, incorporating character-level information during the alignment process could help address the challenges posed by di erent scripts and transliteration variations between English and Hindi. Finally, implementing attention mechanisms within the adversarial framework could enable more fine-grained mappings that adapt to di erent semantic contexts, potentially addressing current limitations in translating culturally specific concepts or idioms that lack direct equivalents across languages

5 Bibiliography

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings." arXiv preprint arXiv:1805.06297 (2018).

Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the association for computational linguistics 5 (2017): 135-146.

Conneau, Alexis, et al. "Word translation without parallel data." arXiv preprint arXiv:1710.04087 (2017).

Athiwaratkun, Ben, Andrew Gordon Wilson, and Anima Anandkumar. "Probabilistic fasttext for multi-sense word embeddings." arXiv preprint arXiv:1806.02901 (2018).