Cross-Lingual Word Embedding Alignment

Background

Cross-lingual word embeddings are crucial for various multilingual NLP tasks. This assignment focuses on aligning monolingual word embeddings from English and Hindi to create a shared cross-lingual embedding space.

Assignment Question

Implement and evaluate a supervised cross-lingual word embedding alignment system for English and Hindi using the Procrustes method. Follow the steps below to complete this task.

Steps

1. **Data Preparation:**
   a. Train monolingual FastText word embeddings for English and Hindi using appropriate corpora. If computational resources are limited, you may use pre-trained FastText embeddings. If training your own embeddings, you 10,000 articles from Wikipedia for each language.
   b. Limit your vocabulary to the top 100,000 most frequent words in each language.
   c. Extract a list of word translation pairs from the MUSE dataset to use as a bilingual lexicon for supervised alignment.

2. **Embedding Alignment:**
   a. Implement the Procrustes alignment method to learn a linear mapping between the source (English) and target (Hindi) embeddings using the bilingual lexicon.
   b. Ensure that the mapping is orthogonal to preserve distances and angles between word vectors.

3. **Evaluation:**
   a. Perform word translation from English to Hindi using the aligned embeddings.
   b. Evaluate the translation accuracy using the MUSE test dictionary.
   c. Report Precision@1 and Precision@5 metrics for the word translation task.
   d. Compute and analyze cosine similarities between word pairs to assess cross-lingual semantic similarity.
   e. Conduct an ablation study to assess the impact of bilingual lexicon size on alignment quality. Experiment with different training dictionary sizes (e.g., 5k, 10k, 20k word pairs).

Submission Requirements

Source code (Python, preferably a colab notebook) with clear documentation and instructions for reproducing your results.

Optional Extra Credit

Implement an unsupervised alignment method such as Cross-Domain Similarity Local Scaling (CSLS) combined with adversarial training, as described in the MUSE paper. Compare its performance with the supervised Procrustes method.

Resources

- MUSE dataset and pre-trained embeddings: https://github.com/facebookresearch/MUSE
- FastText: https://fasttext.cc/
- Procrustes alignment method: Described in "Word Translation Without Parallel Data" by Conneau et al. (2017)