An Exploration in L2 Word Embedding Alignment

Peiyu Liao
pyliao@stanford.edu
Department of Computer Science

The main goal of this project is to verify the hypothesis that word embeddings trained from an L2 corpus better align with the source L1 embeddings than the target L1 embeddings. If the hypothesis can be verified, better unsupervised cross-lingual word alignment can be achieved, which helps in some downstream tasks such as Unsupervised MT.

Dataset (# tokens)
Source L1 Corpus (Chinese, by Chinese speakers)
Wikipedia (54M), Wikipedia (topics match with arXiv categories, 14.3M), Weibo (18.9M)
Target L2 Corpus (English, by Chinese speakers)
arXiv (14.3M), English learner essay corpus (14.8M)
Target L1 Corpus (English, by English speakers)
arXiv (8.6M), Twitter (27B), Common Crawl (600B)

Approach
Word Embedding Training
word2vec (w), fastText (f), GloVe (g)

Word Embedding Space Alignment
Learn a linear mapping $W$ by solving:

$$W^* = \arg \min_{W \in M_d(\mathbb{R})} \|WX - Y\|_F$$

Employ adversarial training to learn without parallel data $X, Y$.

Analysis

Hypothesis Verification
Compare L1 and L2 experiments, there is no clear winner in both supervised and unsupervised alignment. Hypothesis remains to be verified.

Conclusion
The hypothesis that L2 corpus can achieve better alignment is not verified from the experiments. However, this project demonstrates a new use of L2 corpus in training word embeddings.

Key findings were found to help further improvement in unsupervised word alignment, including the benefits of a larger vocabulary size and a naturally-aligned corpus.

References