

# An Exploration in L2 Word Embedding Alignment

Peiyu Liao

pyliao@stanford.edu

Department of Computer Science

## Problem

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)

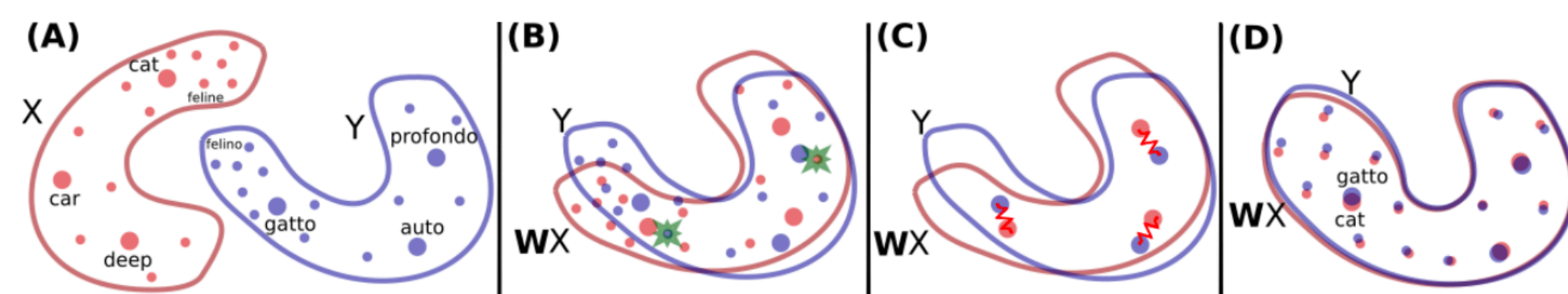
arXiv (8.6M), Twitter (27B), Common Crawl (600B)

## 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  $X, Y$  [1].

## Word Translation Precision@k Results

source-target	# Sent (tgt)	Emb	Word Sim (tgt)	NN p@1	NN p@5	NN p@10	CSLS p@1	CSLS p@5	CSLS p@10
wiki-wiki (full)	-	f	0.65	32.93 / -	50.53 / -	57.52 / -	36.64 / -	55.95 / -	62.22 / -
wiki-wiki	-	f	0.65	1.21 / -	3.42 / -	4.56 / -	1.07 / -	4.2 / -	6.70 / -
wiki-crawl	-	f	0.71	0 / -	0 / -	0 / -	0 / -	0.075 / -	0.075 / -
wiki_t-arxiv_l2	880K	w	0.24	0 / 4.55	0 / 10.88	0.14 / 14.74	0 / 5.37	0 / 11.16	0.14 / 15.01
wiki_t-arxiv_l2	480K	w	0.21	0 / 3.31	0 / 8.12	0.15 / 10.53	0 / 3.61	0 / 8.57	0 / 10.98
wiki_t-arxiv_l1	480K	w	0.175	0 / 3.46	0 / 7.59	0 / 10.12	0 / 3.86	0 / 7.19	0 / 10.25
wiki_t-arxiv_l2	880K	f	0.072	0 / 2.07	0 / 4.82	0.14 / 6.47	0 / 2.20	0 / 4.27	0.14 / 6.47
wiki_t-arxiv_l2	480K	f	0.031	0 / 1.80	0 / 3.46	0.15 / 5.86	0 / 1.95	0 / 3.91	0 / 5.86
wiki_t-arxiv_l1	480K	f	0.175	0 / 1.46	0 / 4.39	0 / 6.39	0 / 1.86	0 / 4.39	0 / 6.13

\* **wiki-wiki (full)**: original paper setting, larger max vocabulary size

\* **wiki\_t**: Wikipedia of certain topics; **arxiv\_lx**: arXiv L1 or L2 corpus

\* **x / y** in precisions: x for unsupervised results, y for supervised results

\* **NN**: nearest neighbor; **CSLS**: a similarity measurement defined in [1]

\* Full experimental results please refer to report

## Analysis

The general performance is extremely poor. Possible reasons:

### The Effect of Corpus Size and Vocabulary Size

Compare 880K to 480K experiments, more data is likely beneficial to the alignment performance.

But compare wiki-wiki (full) to wiki-wiki, larger corpus is indeed critical in that it provides larger vocabulary size.

### The Effect of Training Epochs

The performance stops improving within 20 epochs. Training for longer may not be helpful.

### The Effect of Corpus Choice

**Key factor!** wiki-crawl does not show promising results even though it is trained from a large corpus. Wikipedia corpus is naturally aligned and is the most ideal corpus for this task, regardless of whether our hypothesis holds.

### The Effect of Word Embedding Methods

Word2vec is generally better than fastText. The best performance from wiki-wiki (full) may be further improved by using word2vec models.

### 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

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