Precog Recruitment Task — Language Representations

Part 2: Cross-lingual Alignment

Nilanjan Sarkar
MS by Research, CSE
ID: 2025701014

September 29, 2025

1 Problem Overview

The goal is to align independently trained monolingual word embeddings for English and Hindi so that words with the same meaning are close across languages. We use fastText vectors and a small supervised seed dictionary to learn linear mapping functions. Effectiveness is measured with **Bilingual Lexicon Induction (BLI)**—recovering translations by nearest neighbour search—reported as **Precision@k (P@k)** and **Mean Reciprocal Rank (MRR)**. We also run a small qualitative check and a tiny sentence-retrieval toy test.

2 Approach

Data. 300-dimensional fastText embeddings with vocabulary caps of 120,000 each; MUSE-style English–Hindi seed dictionary .
Preprocessing. Keep the top-N words, build word→ index maps, and L2-normalize all vectors.
Mappings. Three systems:

- **NoMap** baseline: no transformation; retrieves translations directly, expected to fail because spaces are misaligned.
- **LLS** (*Linear Least Squares*): $W = \arg\min_W \|XW - Y\|_F^2 = (X^\top X)^{-1} X^\top Y$.
- **Orthogonal Procrustes** (OP): $W^\star = \arg\min_{W:\, W^\top W = I} \|XW - Y\|_F^2$. Let $M = Y^\top X = U\Sigma V^\top$, then $W^\star = VU^\top$.

Retrieval. We compare plain cosine similarity with **Cross-domain Similarity Local Scaling (CSLS)**. CSLS addresses hubness by subtracting average neighbourhood similarities in each space: for source x and target y,

$$\mathrm{CSLS}(x, y) = 2\,\cos(xW, y) - r_S(x) - r_T(y),$$

where $r_S(x)$ (resp. $r_T(y)$) is the average cosine of x (resp. y) to its top-k nearest neighbours in the other space (we use $k = 10$).
Evaluation. Bilingual Lexicon Induction (BLI) using P@1/5/10 and MRR on held-out test pairs.

3 Results

EN→HI BLI

Method	Cosine			CSLS		
	P@1	P@5	P@10	P@1	P@5	P@10
NoMap	0.000	0.000	0.000	0.000	0.000	0.000
LLS	0.115	0.263	0.341	0.205	0.420	0.496
OP	**0.173**	**0.360**	**0.433**	**0.214**	0.413	0.493

MRR (EN→HI): LLS-Cos 0.180; LLS-CSLS 0.295; OP-Cos 0.252; OP-CSLS **0.300**.

Dictionary coverage. 38,221 raw lines; 24,697 in-vocab pairs kept (64.6%). Splits: 19,757/2,469/2,471 (train/dev/test).

4 Discussion

NoMap baseline is near-zero as expected. Both *Linear Least Squares (LLS)* and *Orthogonal Procrustes (OP)* show large gains, and *CSLS* consistently boosts retrieval by reducing hubness. EN→HI shows OP+CSLS as the best by MRR (0.300).