# Discovering Low-rank Subspaces for Language-agnostic Multilingual Representations

Zhihui Xie[1] Handong Zhao[2] Tong Yu[2] Shuai Li[1]

[1]Shanghai Jiao Tong University [2]Adobe Research

## Summary


- We show that there exist **low-rank subspaces** in the pretrained multilingual language models (ML-LMs) that mainly encode **language-specific** signals
- We present a simple approach **LSAR** to identify the subspace in a ML-LM in an **unsupervised** manner (i.e., without any translation pairs)
- Empirical results show that LSAR can remove language-specific signals to **facilitate cross-lingual tasks** that only consider semantic information
- We demonstrate that the subspace encodes **strong syntactic signals** with experimental analysis


## Language-agnostic Representations

- ML-LMs like mBERT and XLM-R exhibit **impressive cross-lingual ability**
- But previous works observe that these ML-LMs encode **strong language identity information**
- Key question:

*"Can we extract the language-agnostic part to benefit tasks that only consider semantic information?"*

- It is often assumed that each embedding $\boldsymbol{e}_l$ in language $l$ can be decomposed in an additive form:

$$\boldsymbol{e}_l := \boldsymbol{s}_l + \boldsymbol{a}_l$$

## Low-rank Subspaces in ML-LMs

Our method LSAR is simple but effective

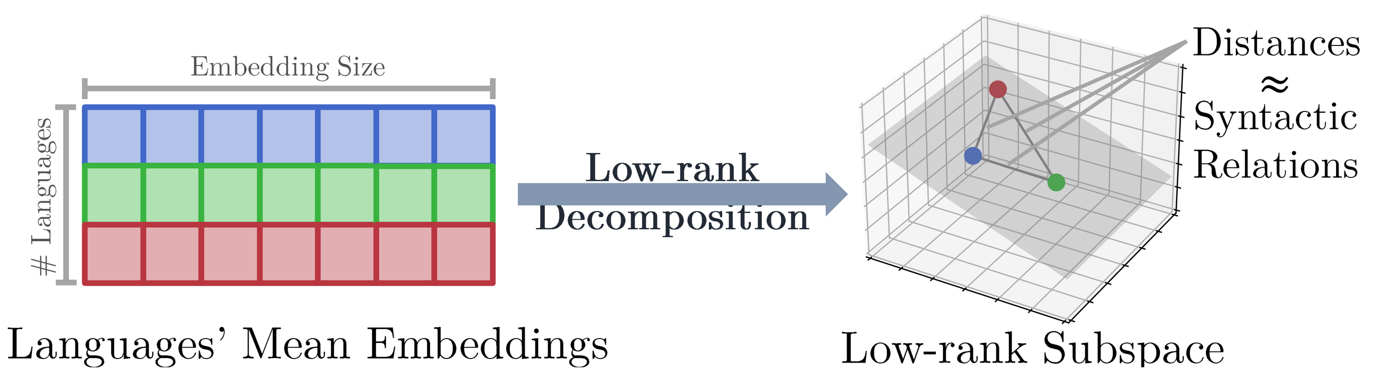



Figure 1. A conceptual illustration of our alignment method LSAR.

- Extract $d$-dimensional embeddings from monolingual corpora (e.g., OSCAR) of $L$ languages using the ML-LM to obtain a mean embedding matrix $\boldsymbol{M} \in \mathbb{R}^{d\times L}$
- Decompose $\boldsymbol{M}$ into two components: a vector $\boldsymbol{\mu} \in \mathbb{R}^d$ shared among languages and a matrix $\boldsymbol{M}_s \in \mathbb{R}^{d\times r}$ representing a low-rank subspace on which linguistic signals are expressed differently for each language:

$$\min_{\boldsymbol{\mu},\boldsymbol{M}_s,\boldsymbol{\Gamma}} \left\|\boldsymbol{M} - \boldsymbol{\mu}\mathbb{1}^\top - \boldsymbol{M}_s\boldsymbol{\Gamma}^\top\right\|_F^2$$
$$\text{s.t. } \boldsymbol{\mu} \perp \operatorname{Span}(\boldsymbol{M}_s)$$

- Project embeddings onto the null space of $\boldsymbol{M}_s$:

$$\begin{aligned}\boldsymbol{a}_l &= \left(\boldsymbol{I} - \boldsymbol{M}_s\left(\boldsymbol{M}_s^\top\boldsymbol{M}_s\right)^{-1}\boldsymbol{M}_s^\top\right)\boldsymbol{e}_l \\ &= \boldsymbol{e}_l - \boldsymbol{M}_s\boldsymbol{M}_s^\top\boldsymbol{e}_l\end{aligned}$$

## Experimental Results

Applying LSAR **consistently leads to improvements** over commonly used ML-LMs

| | mBERT | XLM | XLM-R | LABSE |
|---|---|---|---|---|
| Original | 37.53 | 28.13 | 57.68 | 95.47 |
| Centered | 39.57 | 27.13 | 61.08 | 95.56 |
| LIR ($k=1$) | 39.70 | 28.75 | 61.60 | **95.63** |
| LIR ($k=15$) | 41.21 | 31.65 | 62.80 | 95.56 |
| LSAR | **44.64** | **33.16** | **65.05** | 95.54 |

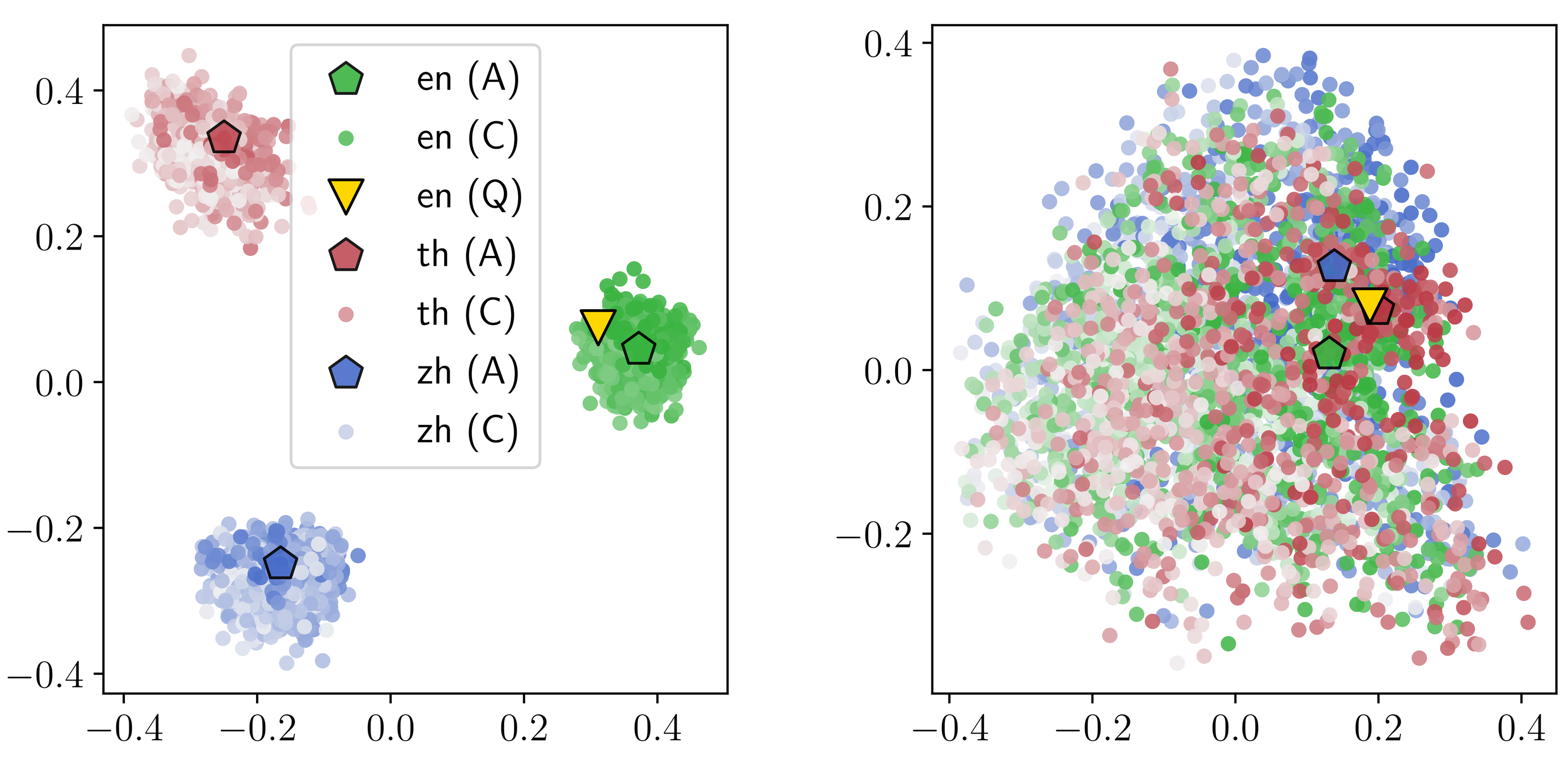
Table 1. Retrieval accuracy (%) on **Tatoeba** (averaged over all 36 languages).

| | XQuAD-R | | MLQA-R | |
|---|---|---|---|---|
| | En-En | X-X | En-En | X-X |
| Original | 28.57 | 23.36 | 35.71 | 26.21 |
| Centered | 35.37 | 44.66 | 35.36 | 42.14 |
| LIR ($k=1$) | 37.70 | 44.25 | 38.03 | 41.96 |
| LSAR | **41.13** | **45.89** | **40.55** | **43.32** |

Table 2. Answer retrieval mAP (%) on XQuAD-R and MLQA-R of **LAReQA** (averaged over all languages).

## Analysis

- LSAR effectively removes **same-language bias**




(a) Original (b) LSAR

Figure 2. **2D PCA visualization** on LAReQA. We display the embeddings collected from mBERT (X-X) on the XQuAD-R sub-dataset. Embeddings of the candidate answers (C) in English, Thai, and Mandarin are shown in small scatters. Embeddings of the question (Q) in English and the ground-truth answers (A) in English, Thai, and Mandarin are shown in large scatters.

- The subspace primarily encodes **syntactic information**

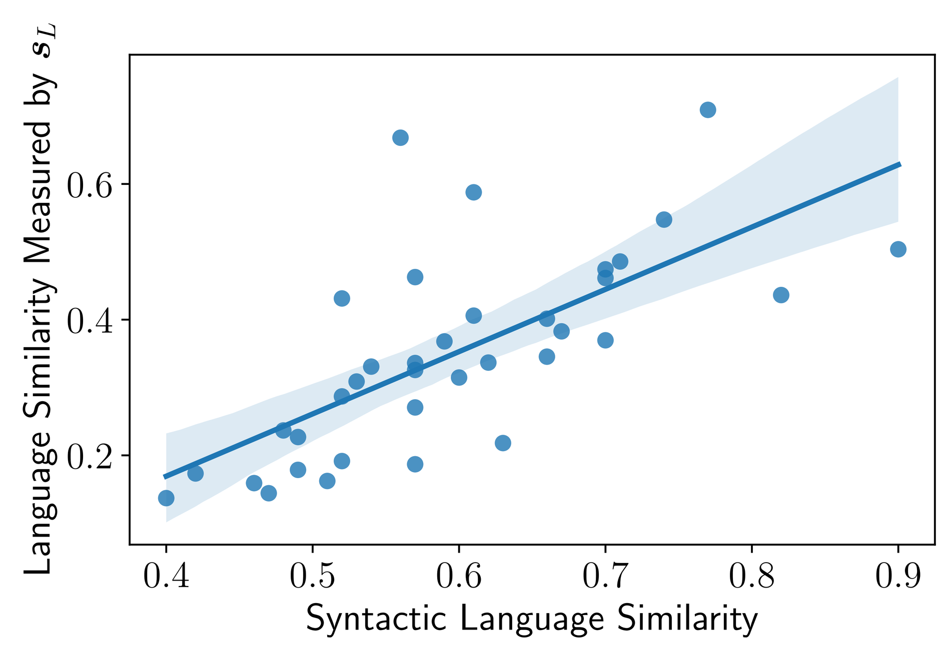



Figure 3. Language similarity obtained from **syntactic signals vs. language similarity** measured by language-specific $\boldsymbol{s}_L$ of mBERT. Each point is a language.