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

{fffffarmer,shuaili8}@sjtu.edu.cn

{hazhao,tyu}@adobe.com

Abstract

Large pretrained multilingual language models (ML-LMs) have shown remarkable capabilities of zero-shot cross-lingual transfer, without direct cross-lingual supervision. While these results are promising, follow-up works found that, within the multilingual embedding spaces, there exists strong language identity information which hinders the expression of linguistic factors shared across languages. For semantic tasks like cross-lingual sentence retrieval, it is desired to remove such language identity signals to fully leverage semantic information. In this work, we provide a novel view of projecting away language-specific factors from a multilingual embedding space. Specifically, we discover that there exists a low-rank subspace that primarily encodes information irrelevant to semantics (e.g., syntactic information). To identify this subspace, we present a simple but effective *unsupervised* method based on singular value decomposition with multiple monolingual corpora as input. Once the subspace is found, we can directly project the original embeddings into the null space to boost language agnosticism without finetuning. We systematically evaluate our method on various tasks including the challenging language-agnostic QA retrieval task. Empirical results show that applying our method consistently leads to improvements over commonly used ML-LMs.

1 Introduction

Large language models pretrained with self-supervised objectives (e.g., masked language modeling) have become the *de-facto* standard for various NLP tasks (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019). Follow-up extensions to the multilingual setting inherit similar training objectives and show very promising results (Conneau and Lample, 2019; Conneau et al., 2020b; K et al., 2020). Despite these models are trained

---
[*]Corresponding author.

without explicit cross-lingual signals (i.e., translation pairs), they surprisingly exhibit impressive zero-shot cross-lingual transferability on natural language inference (Conneau et al., 2018), question answering (Lewis et al., 2020), sentence retrieval (Artetxe and Schwenk, 2019), etc.

While these ML-LMs offer practical solutions for cross-lingual tasks, there is an enduring debate about why the ML-LMs work. From a positive perspective, Pires et al. (2019) conduct an exploratory study on mBERT (Devlin et al., 2019), suggesting that cross-lingual transfer is possible even to languages in different scripts. Chi et al. (2020) probe mBERT for structural phenomena and find that its representations can recover syntactic tree distances in languages other than English. These findings present shreds of evidence that the pretrained multilingual representations do capture cross-lingual properties in various aspects. On the flip side, a line of research shows that pretrained ML-LMs encode strong *language-specific* signals. This causes their multilingual representations to cluster by language identities instead of semantic meaning (Wu and Dredze, 2019; Roy et al., 2020; Libovický et al., 2020). The property largely hinders the expression of linguistic signals shared across languages. For applications like cross-lingual sentence retrieval that mainly consider semantic information, ML-LMs with strong language-specific signals tend to retrieve answers from specific languages, regardless of their semantic meaning (Roy et al., 2020).

Motivated by previous findings about language identity information, we aim to locate language-specific factors captured by the pretrained ML-LMs for recovering a *language-agnostic* embedding space. Inspired by advances in domain generalization (Muandet et al., 2013; Motiian et al., 2017; Piratla et al., 2020), we explore a simple but effective approach, LSAR, to discover a **L**ow-rank **S**ubspace for language-**A**gnostic **R**epresentations within an ML-LM. The subspace primarily encodes

information irrelevant to semantics, and can be identified *without any translation pairs* based on singular value decomposition. Once the subspace is found, we can directly factor out language-specific factors from the multilingual embeddings by projecting them into the null space without finetuning.

To evaluate LSAR, we focus on semantic tasks for multilingual sentence embeddings. On standard cross-lingual zero-shot transfer tasks including classification and sentence retrieval, LSAR consistently achieves significant improvements. Especially, applying LSAR leads to significant improvements for pretrained ML-LMs on LAReQA (Roy et al., 2020), a challenging benchmark targeting strong language agnosticism.

We further examine what information exactly the subspace contains. By performing correlation analysis between structural language similarities obtained from the URIEL database (Littell et al., 2017) and the language similarities captured on the subspace, we observe that the subspace encodes a great deal of syntactic information. This implies that LSAR successfully erases linguistic signals that are redundant to semantic tasks to facilitate language agnosticism.

To conclude, our main contributions are:

- We present one of the pioneering efforts to discover that there exist low-rank subspaces of pretrained ML-LMs' embeddings that mainly encode language-specific signals.

- To identify the subspace in a ML-LM, we present a simple unsupervised approach called LSAR based on singular value decomposition. By projecting embeddings onto the null space, LSAR can exclude the unwanted factors to facilitate language agnosticism.

- Empirical results show that LSAR is surprisingly effective for a variety of semantic tasks. We also elucidate that the subspace encodes strong syntactic signals with careful experimental analysis.

2 Related Work

Understanding Pretrained Multilingual Representations   Recently, there has been a surge of interest in probing pretrained ML-LMs like mBERT (Devlin et al., 2019). Pires et al. (2019) present an exploratory study on the cross-linguality of mBERT, showing that mBERT exhibits strong

zero-shot performances for typologically similar languages. Libovický et al. (2020) find that the original mBERT embeddings can be decomposed into a language-specific component and a language-neutral component. Chi et al. (2020) probe mBERT for universal grammatical relations and show that mBERT does encode fine-grained syntactic distinctions across languages. Muller et al. (2021) find that mBERT operates as the stacking of two subnetworks and mainly the lower part of the model is crucial for cross-lingual transfer.

Language-agnostic Representations   To further facilitate semantic downstream tasks like text classification, retrieval, and question answering, it is appealing to remove language-specific signals from the original embeddings without destroying the intrinsic semantic meaning.

LASER (Artetxe and Schwenk, 2019) utilizes parallel data to train a BiLSTM-based multilingual sentence encoder. Zhao et al. (2021) obtain language-agnostic embeddings from mBERT and XLM-R by explicitly aligning the word pairs and further normalizing the latent spaces with zero mean and unit variance. Yang et al. (2021) regard the top principal components from each language's embedding space as the primary source of language bias and propose to project them away to boost language agnosticism.

Our work bears resemblance to Yang et al. (2021), but with clear distinctions in that: 1) we model language-specific signals *jointly* in the multilingual embedding space instead of locating it *separately* within each language; 2) we further verify what exactly the linguistic signals are identified, and present evidences that LSAR primarily removes syntactic information. A few previous works (Gonen et al., 2020; Liang et al., 2021; Chang et al., 2022) also attempt to locate language-agnostic embeddings in subspaces of ML-LMs. Apart from the dissimilarity of methodology, we focus on sentence-level instead of token-level tasks and provide shreds of evidence that the identified subspace exhibits strong correlation with syntactic information.

Low-rank Subspaces in Other Applications   Low-rank subspaces have been employed in various applications.   In face recognition, the most expressive features for face representations are located via subspace analysis methods like PCA (Turk and Pentland, 1991; Wang and Tang,
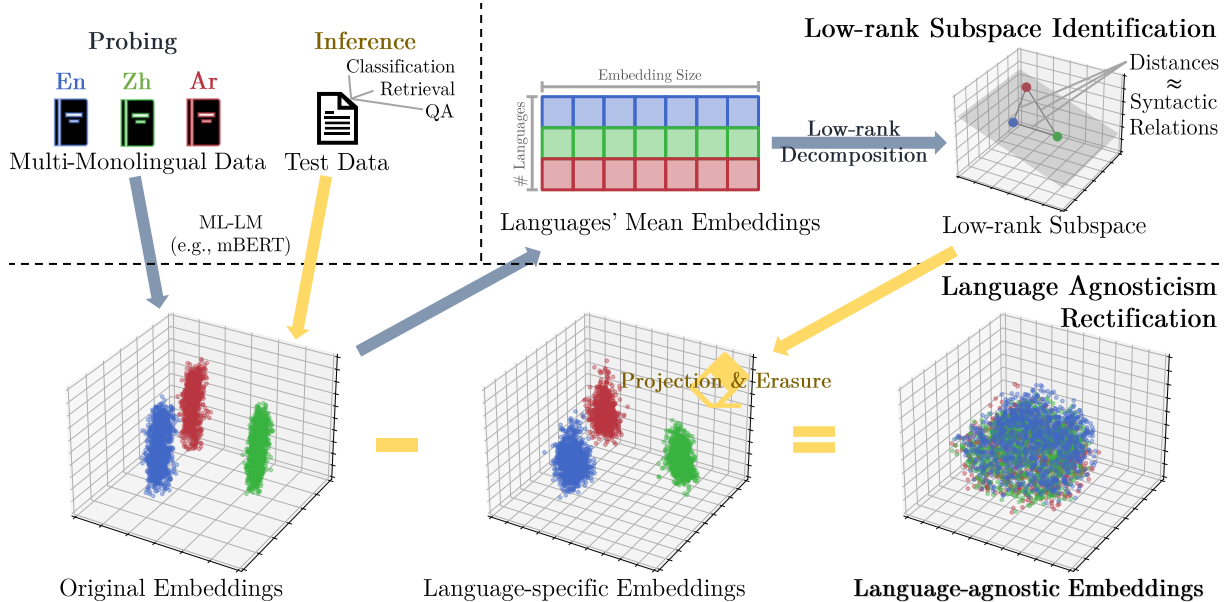
Figure 1: Conceptual illustration of our alignment method LSAR. There exists strong language identity information from the original pretrained multilingual representations. By projecting away language-specific components that reside in a low-rank subspace discovered in identification process (in top-right), we can produce a language-agnostic embedding space via language agnosticism rectification (in bottom). The probing procedure (colored in blue-grey) and the inference procedure (colored in yellow) can be done separately.

2004). For domain adaptation and domain generalization, a typical idea is to uncover a shared subspace on which the distribution mismatch between domains is reduced (Muandet et al., 2013; Pan et al., 2011; Motiian et al., 2017). Recent advances in probing Generative Adversarial Networks (GANs) also observe meaningful latent subspaces that enable precise control of GAN generation (Wang and Ponce, 2021; Zhu et al., 2021). These findings to some extent motivate this paper.

3 Methodology

In this section, we first introduce our method to identify the low-rank language-specific subspace in an *unsupervised* manner. Once the subspace is found, we can then suppress the language identity from the original multilingual embeddings to achieve language agnosticism rectification by projecting them to the null space. This post-training alignment procedure can largely benefit downstream tasks like cross-lingual retrieval which solely utilize semantic-related information.

3.1 Multilingual Embedding Decomposition

To locate the language-specific factors, we follow previous works (Pires et al., 2019; Libovický et al., 2020; Yang et al., 2021) to hypothesize that each multilingual embedding $e_l \in \mathbb{R}^d$ in language l can

be decomposed in an additive form:

$$e_l := s_l + a_l,$$

where $s_l \in \mathbb{R}^d$ and $a_l \in \mathbb{R}^d$ represent the language-specific component to remove and the language-agnostic component to keep, respectively.

Built on the above assumption, previous unsupervised approaches extract the language identity information *separately* for each language space. Given an ML-LM (e.g., mBERT), the extracted embeddings $\mathcal{E}_l := \{e_l^i\}_{i=1}^n$ from n samples of task training data or external monolingual corpora contain mixed linguistic information of semantic-relevant and semantic-irrelevant signals about language l. Libovický et al. (2020) use the empirical mean $\frac{1}{n}\sum_{i=1}^n e_l^i$ to obtain s_l. Yang et al. (2021) use the top-k principal components $C_l = \text{PCA}(\mathcal{E}_l) \in \mathbb{R}^{d \times k}$ to encode language identity signals, and propose to factor them out with $s_l = C_l C_l^\top e_l$ to facilitate language agnosticism.

In spite of their promising results for semantic-related tasks, these methods fall short of comprehensively discovering cross-lingual relationship in the latent space. For each language l, both of them leverage solely $\mathcal{E}_l$ to locate language-specific information, which fails to distinguish itself from semantic signals as other languages' characteristics is unknown. Without careful tuning, this can

lead to unexpected semantic information loss (Khodak et al., 2018). Besides, it is also unclear what exactly language-specific signals are captured by these approaches.

3.2 Low-rank Subspace Identification

To alleviate the above issues, we attempt to globally capture language-specific information from the multilingual latent space. Inspired by previous works in domain adaptation and domain generalization (Muandet et al., 2013; Motiian et al., 2017; Piratla et al., 2020), we present a simple approach that identifies a low-rank subspace of the original multilingual latent space, $M_s \in \mathbb{R}^{d \times r}$, spanned by r components. Intuitively, the subspace encodes language-specific signals via measuring the latent discrepancy among languages.

To be specific, we first extract the mean embedding $\boldsymbol{\mu}_l = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{e}_l^i$ of each language l in the same spirit of previous approaches. Concatenating $\boldsymbol{\mu}_l$ of L languages column-by-column results in the mean embedding matrix $M \in \mathbb{R}^{d \times L}$. As discussed in Section 3.1, the mean embeddings can unexpectedly mix the desired language-specific signals with semantic information. To avoid removing the semantic information shared among languages, we decompose M into two components: 1) a vector $\boldsymbol{\mu}$ representing what is commonly shared across languages in the latent space; 2) a matrix M_s specifying a low-rank subspace on which different languages express different linguistic signals. With the orthogonality constraint, our objective is:

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

where $\boldsymbol{\Gamma} \in \mathbb{R}^{L \times r}$ is the coordinates of language-specific signals along the subspace's r components and $\mathbb{1} \in \mathbb{R}^d$ contains all ones.

The optimal solution of Equation 1 can be computed efficiently via Singular Value Decomposition (SVD), as proved in Appendix A. Algorithm 1 presents the detailed procedure. The only hyperparameter $r < L$ controls the amount of language-specific information captured by the identified subspace. The larger r is, the more language-specific signals we can identify.

3.3 Language Agnosticism Rectification

Once we find the low-rank subspace with semantically irrelevant information encoded, we can im-

---

Algorithm 1: language-specific Subspace Identification

In: languages' mean embeddings M, rank of subspace r
Out: language-agnostic component $\boldsymbol{\mu}$, language-specific subspace M_s, coordinates $\boldsymbol{\Gamma}$

/* 1) Approximate M in low rank */
1 $\boldsymbol{\mu}' \leftarrow \frac{1}{d} M \mathbb{1}$;
2 $M_s', {}_-, \boldsymbol{\Gamma}' \leftarrow \text{Top-}r \text{ SVD} \left(M - \boldsymbol{\mu}' \mathbb{1}^\top \right)$;
3 $M' \leftarrow \boldsymbol{\mu}' \mathbb{1}^\top + M_s' \boldsymbol{\Gamma}'^\top$;
   /* 2) Force orthogonality   */
4 $\boldsymbol{\mu} \leftarrow \frac{1}{\| M'^+ \mathbb{1} \|^2} M'^+ \mathbb{1}$;
5 $M_s, {}_-, \boldsymbol{\Gamma} \leftarrow \text{Top-}r \text{ SVD} \left(M' - \boldsymbol{\mu} \mathbb{1}^\top \right)$

---

prove language agnosticism via projecting multilingual embeddings onto the null space of M_s:

$$\boldsymbol{a}_l = \left(I - M_s \left(M_s^\top M_s \right)^{-1} M_s^\top \right) \boldsymbol{e}_l$$
$$= \boldsymbol{e}_l - M_s M_s^\top \boldsymbol{e}_l.$$

Given that usually $l \ll d$, the information removed is restricted to aspects that emerges to be language-specific and will not lead to dimensional collapse.

4 Experiments

We systematically evaluate our method on various tasks followed by further analyses[1], with the purposes of understanding: 1) whether the proposed approach can benefit downstream tasks; 2) what exactly the identified low-rank subspace captures.

To begin with, we describe our evaluation protocol for the alignment methods, which largely follows Yang et al. (2021) but with a broader scope to include more base models as listed in Section B. Given one of the pretrained ML-LMs, we first randomly collect 10,000 sentences for each language from the OSCAR corpus (Ortiz Suárez et al., 2020) covering all the evaluated languages and their web crawl texts[2]. The sentence embeddings extracted by the pretrained model are then used for finding

---

[1]Code: https://github.com/fffffarmer/LSAR.
[2]Yang et al. (2021) use Wiki-40B (Guo et al., 2020) for collecting sentence embeddings. The corpus fails to cover all the languages evaluated in Tatoeba. We also report the numbers using Wiki-40B as the text resource for LAReQA and Amazon Reviews in Appendix C.2.

	mBERT	XLM	XLM-R	LABSE
Cross-lingual zero-shot transfer (w/o finetuning)				
Original	37.53+00.00%	28.13+00.00%	57.68+00.00%	95.47+00.00%
Centered (Libovický et al., 2020)	39.57+05.43%	27.13-03.57%	61.08+05.89%	95.56+00.10%
LIR ($k = 1$) (Yang et al., 2021)	39.70+05.77%	28.75+02.22%	61.60+06.80%	**95.63**+00.16%
LIR ($k = 15$) (Yang et al., 2021)	41.21+09.80%	31.65+12.51%	62.80+08.87%	95.56+00.10%
LSAR	**44.64**+18.94%	**33.16**+17.89%	**65.05**+12.77%	95.54+00.08%
Cross-lingual zero-shot transfer (w/ finetuning)				
Full-Model-FS (Xu et al., 2022)[†]	-	-	60.5+04.9%/66.2+14.8%	-
S[4]-Tuning (Xu et al., 2022)[†]	-	-	66.1+14.6%/69.5+20.5%	-
Full-Model (Ruder et al., 2021)[‡]	42.8+14.0%	-	76.6+32.8%	-

Table 1: Retrieval accuracy (%) on Tatoeba (averaged over all 36 languages). [†]Results from Xu et al. (2022) report few-shot performances with different numbers of shots (64/128). [‡]Results are calculated from Ruder et al. (2021). We use "-" to indicate results that are not reported in the references and use "+%" to report relative improvements.

the low-rank subspace described in Equation 1. Unless otherwise indicated, we consistently report LSAR with $r = l - 1$, where l is the number of the evaluated languages. We evaluate language agnosticism over pretrained ML-LMs that are commonly used, as described in Appendix B. Detailed results are listed in Appendix C.3.

4.1 Baselines

Apart from **Original** that keeps the pretrained ML-LM intact, we compare LSAR with the following baselines. The baselines share the same setting as ours in that both of them require no parallel text and aim at removing language-specific factors in a post-training manner.

Centered Libovický et al. (2020) extract language-neutral embeddings from the original pretrained multilingual sentence encoders via subtracting the mean embedding for each language. The mean embeddings are calculated from the multi-monolingual OSCAR corpus.

LIR Yang et al. (2021) propose to project away the top-k principal components of each language's embeddings to facilitate language agnosticism, where k is the hyperparameter. Again, the top principal components are extracted from the multi-monolingual corpus.

4.2 Sentence Retrieval

Tatoeba (Artetxe and Schwenk, 2019) is a commonly used dataset for evaluating ML-LMs. It comprises up to 1,000 sentences for each language along with their English translations. We follow the evaluation procedure of XTREME (Hu et al.,

2020) that covers 36 languages. For each language pair, we go through each sentence in the source language and find the closest sentence in the target language using cosine similarity.

The top-1 retrieval accuracy results are shown in Table 1. For mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), and XLM-R (Conneau et al., 2020a), applying LSAR brings significant performance gains of up to 19% relative improvement. Compared with Centered and LIR which separately remove information for each language, LSAR jointly utilizes the encoded information from all the languages to better locate language-specific factors. Furthermore, we observe that LSAR consistently achieves the best results with hyperparameter r (the rank of the low-rank subspace) equal to the number of the evaluated languages, as shown in Appendix C.1. As the languages are diversely distributed, it is reasonable that each language possesses its own linguistic characteristics, resulting in a larger language-specific subspace to factor out. In contrast, we find that LIR is vulnerable to its hyperparameter k (the number of the removed principal components), which is best shown in Figure 7.

For LABSE (Feng et al., 2022), all the methods fail to provide marked enhancement. This can be mainly attributed to the fact that LABSE already uses parallel corpora to explicitly align multilingual embeddings. Despite that the improvement is marginal, it is still promising to combine LSAR with existing pretraining objectives to produce better language-agnostic embeddings.

We also include several representative baselines that finetune either mBERT or XLM-R for bet-

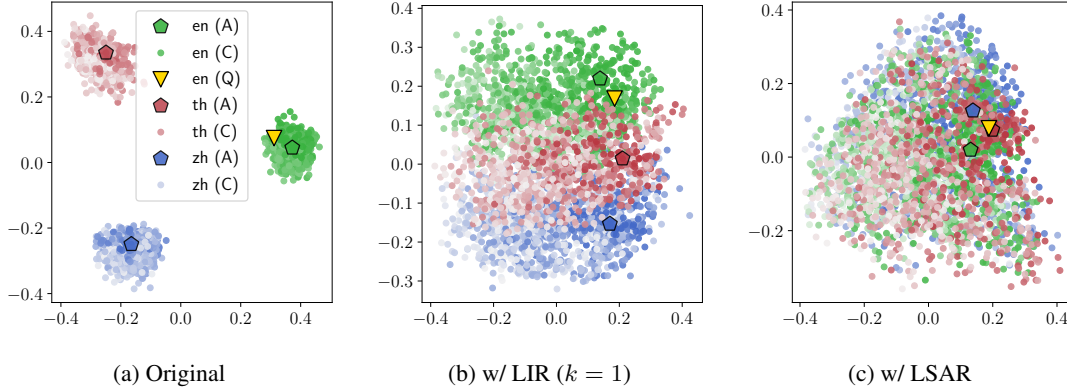(a) Original   (b) w/ LIR ($k = 1$)   (c) w/ LSAR

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. Higher opacity indicates higher predicted ranking (color bars: ▬/▬/▬).

	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).

ter cross-lingual transfer results. Although these methods are not directly comparable to ours, we believe it provides additional valuable findings to include them. Full-Model-FS and S⁴-Tuning finetune XLM-R on full English labeled examples and then K-shot data over target languages ($K = 64/128$). For Full-Model, the pretrained models are finetuned on the English SQuAD data. On mBERT, LSAR outperforms Full-Model by a large margin. We also observe on XLM-R that LSAR is competitive with finetuning the full model on 128-shot data as well as finetuning a dedicated language subnetwork (S⁴-Tuning) on 64-shot data. The results are quite promising given that we obtain better performances with the original encoders intact and no task-relevant training data.

4.3 Language-agonstic Answer Retrieval

While Tatoeba reveals the cross-lingual transferability across English-centric language pairs, it is restricted to monolingual pools (i.e., the set of candidates is restricted to certain language). Therefore, it fails to thoroughly evaluate whether texts with a similar semantic meaning are grouped together in

the latent space, regardless of their languages.

With that in mind, we further examine the alignment methods on LAReQA (Roy et al., 2020), a challenging cross-lingual answer retrieval task. Unlike Tatoeba, the targets of LAReQA must be retrieved from a large multilingual candidate pool. It consists of two sub-datasets, XQuAD-R and MLQA-R, whose candidate pool covers 11 and 7 languages respectively.

We follow Yang et al. (2021) to evaluate the alignment methods on two models, **mBERT (En-En)** and **mBERT (X-X)**. Specifically, mBERT (En-En) finetunes the original mBERT model on the English QA pairs collected from the SQuAD v1.1 dataset. mBERT (X-X) employs the same training procedure but with an extended dataset where each sample is translated into the 11 XQuAD languages. Since all positive samples for finetuning are within the same language as the question query, both models exhibit strong self-language bias while preserving the weak alignment property. For evaluation, we use the dot product of embeddings to score a QA pair, which accords with the finetuning protocol. The retrieval performance is measured by mean Average Precision (mAP).

Table 2 reports our LAReQA results. We can observe that applying LSAR again results in signification improvements, nearly doubling mAP of mBERT (X-X) on XQuAD-R. Since in the candidate pool each language has one of the relevant answers, better retrieval performances directly indicate better language agnosticism. Centered and LIR ($k = 1$) also show impressive performances, suggesting that in weakly aligned multilingual sys-

	mBERT	XLM	XLM-R
Original	74.73	75.31	80.32
LIR ($k = 1$)	75.39	75.73	81.14
LSAR ($r = 1$)	**75.58**	74.93	81.47
LSAR ($r = 2$)	75.49	**75.85**	**82.37**
LSAR	75.24	75.27	81.25

Table 3: Classification accuracy (%) on Amazon Reviews (averaged over English, French, German and Japanese). We exclude Centered as the embeddings are already normalized and hence Centered produces the same results as Original. The results of LABSE are placed in Appendix C due to limited space.

tems, the mean embeddings and principal components do encode language-specific signals. But for LIR, it is shown that removing the first principal component consistently leads to the best performance. This is opposite to what we observe on Tatoeba, where the optimal k is around 15.

To further illustrate the degree of language agnosticism, we project an English question (*What theory best explains gravity?*) as well as all candidates and the ground-truth answers in English, Thai, and Mandarin via PCA. As plotted in Figure 2, candidates in English are retrieved from mBERT (X-X) with higher priority than those in Thai and Mandarin. Applying LSAR can effectively eliminate strong language identity information residing in the original embedding space and draw closer the question and answers from different languages. LIR with $k = 1$, however, falls short of rectifying language-specific signals as illustrated by the embedding spectrum in Figure 2b.

4.4 Zero-shot Classification

We also include the Amazon Reviews classification task (Prettenhofer and Stein, 2010) to assess zero-shot cross-lingual transfer. The dataset consists of product reviews in English, French, German, and Japanese. Each review is labeled as positive or negative, making it a binary classification task. We use

	mBERT	XLM	XLM-R
Original	0.2815	0.5422	0.2457
Centered	0.0975	0.2483	0.2004
LIR ($k = 1$)	0.0900	0.1875	0.2203
LSAR	**0.0801**	**0.1320**	**0.0856**

Table 4: Clustering performance (NMI) of embeddings obtained by mBERT on Tatoeba. The results of LABSE are placed in Appendix C due to limited space.
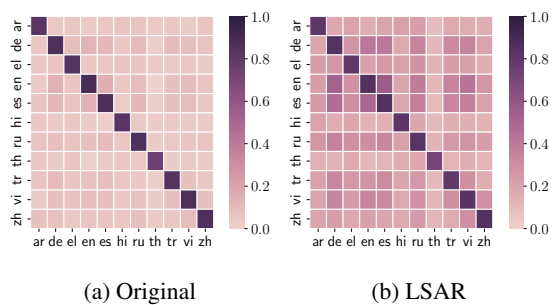

(a) Original (b) LSAR

Figure 3: Answer retrieval mAP on XQuAD-R broken down by question language (row) and answer language (column), with model mBERT (X-X). Only one correct answer is included in the multilingual candidate pool.

the same procedure to extract sentence embeddings as in Section 4.2, and normalize them to make regularization hyperparameters more consistent across languages. Appendix C.1 specifies how we select hyperparameters. Following (Yang et al., 2021), the performance is evaluated via training a logistic regression classifier[3] on the English training data and then evaluating it on the test sets of all four languages.

From Table 3, we observe that the classifier trained on English data benefits from LSAR for classifying reviews based on semantics as the language-specific factors are effectively erased. Another interesting observation is that unlike sentence retrieval, removing more directions does not result in better performance. This indicates that classification tasks can be more sensitive to semantic information.

4.5 Analysis

In this section, we present analysis on a variety of aspects towards what exactly language-specific information LSAR captures.

4.5.1 Language-specific Signals are Rectified

From previous findings, we conjecture that our method achieves impressive cross-lingual performance by effectively removing language identity signals. To quantitatively verify this, we measure the strength of language identity information from the perspective of clustering quality. If the embeddings are clustered by language types, we can generally state that language-specific signals still play a prominent role in the multilingual latent space.

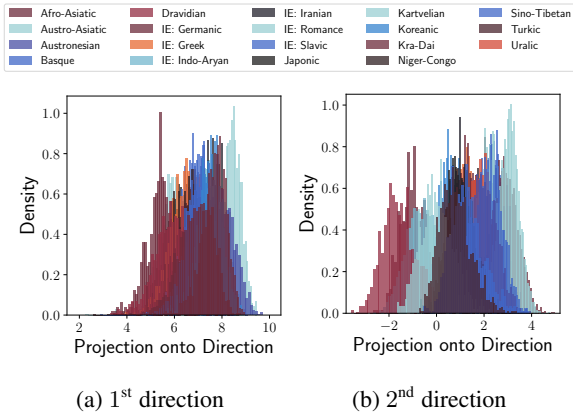[3]sklearn.linear_model.LogisticRegressionCV().

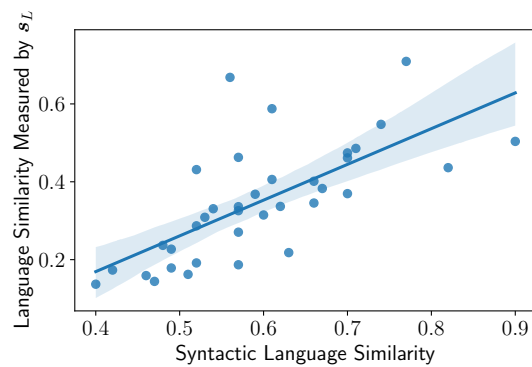Figure 4: Removed components along the top two basis vectors of the identified low-rank subspace on mBERT.

Figure 5: Language similarity obtained from syntactic signals vs. language similarity measured by language-specific s_L of mBERT. Each point is a language.

We perform K-Means clustering on sentence representations of Tatoeba with the number of clusters equal to the number of languages, and then evaluate the resulting clusters using the Normalized Mutual Information (NMI) metric (Jawahar et al., 2019)[4]. As shown in Table 4, the original pretrained embeddings have relatively high NMI scores, suggesting the existence of strong language identity information. Our method consistently achieves smaller NMI scores. This indicates that the embeddings have a lower tendency to group by language types since LSAR successfully winnows down language-specific information.

The same conclusion can be drawn from the limit-to-one-target setting of LAReQA (Roy et al., 2020). Specifically, we remove 10 targets from the multilingual pool of XQuAD-R to evaluate on each target separately. We choose the most biased X-X variant as the base model. The heatmaps in Figure 3 show for each question language (row), the retrieval mAP on the pool containing just one target in different answer languages (column). Since X-X has strong self-language bias, Original shows better performance on the diagonal than off-diagonal. After applying LSAR, we observe a significant increase in average off-diagonal performance (23.76% vs. 5.89%), without sacrificing much on-diagonal performance (81.57% vs. 84.57%). This again verifies that applying LSAR effectively removes language-specific information.

4.5.2 Removed Components Form Groups of Language Families

We next examine whether the removed components found by the low-rank subspace are truly language-

specific. This is demonstrated via plotting the removed components for different languages along top basis vectors of the subspace. For the ease of visualization, we group them by language family.

Figure 4 shows the histograms of removed components along the top two basis vectors extracted from mBERT on 36 languages of Tatoeba, according to Equation 1. We can observe that the removed components disperse in groups of language families along these directions. This implies that the identified subspace do capture language-specific signals and hence removing them along the basis vectors can narrow down latent discrepancy.

4.5.3 The Identified Subspace Primarily Encodes Syntactic Information

Finally, given that the removed components are language-specific, we investigate to what extent the low-rank subspace encodes typological relations among languages. Specifically, we use the URIEL database (Littell et al., 2017) to collect distances between English and other languages set out by experts based on certain typological information (e.g., syntax and phonology). We then compare the typological distances with languages similarities obtained from the removed language-specific embeddings s_L as well as the resulting language-agnostic embeddings a_L by calculating the cosine similarity between languages' mean embeddings.

Among all types of typological signals listed in URIEL, we find that the removed language-specific factors are mostly correlated with syntactic information. Table 5 shows the Pearson correlations on English and other 36 languages from Tatoeba. The removed language-specific component s_L is highly correlated with syntactic infor-

---

[4]sklearn.metrics.normalized_mutual_info_score().

	mBERT	XLM	XLM-R	LABSE
s_L	0.6910	0.6378	0.7526	0.6894
a_L	-0.2711	0.2239	0.1338	-0.2362

Table 5: Pearson correlations between syntactic language similarities obtained from the URIEL database, and the language similarities obtained from language-specific s_L as well as language-agnostic a_L.

mation, whereas the correlation is much smaller in the language-agnostic embedding space with s_L removed. This finding is in line with previous works (Chi et al., 2020; Zhao et al., 2021) that observe the pretrained multilingual models encode rich syntactic information.

We find no prominent correlation between the removed components along certain basis vectors of the subspace and typological information. As we do not presuppose any correspondence between basis vectors and linguistic signals, a specific basis vector falls short of individually encoding language-specific information.

5   Conclusion

We present a simple yet effective approach called LSAR to boost language agnosticism for pretrained multilingual encoders. LSAR identifies a low-rank subspace residing in a pretrained model that primarily encodes language-specific signals in an unsupervised manner via singular value decomposition. Once the subspace is discovered, it can be used to efficiently project away the language identity information. Empirical results demonstrate the great effectiveness of LSAR on semantic tasks and shed light on its ability to locate syntactic relations between languages.

Limitations

Our method LSAR is designed and evaluated for semantic tasks. For future work, we are interested in continuing our study for locating more fine-grained linguistic information, which can potentially boost a larger variety of downstream tasks. While the simplicity of the proposed LSAR is appealing, it also opens up directions for future work by generalizing the first-moment mean embeddings to higher-moment statistics and combining with pretraining objectives in more sophisticated ways.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. *CoRR*, abs/2205.10964.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Carl Eckart and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing word-level translations from multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56, Online. Association for Computational Linguistics.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Melbourne, Australia. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7315–7330, Online. Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2021. Locating language-specific information in contextualized embeddings. *CoRR*, abs/2109.08040.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. 2017. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, Atlanta, Georgia, USA. PMLR.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. 2020. Efficient domain generalization via common-specific low-rank decomposition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7728–7738. PMLR.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden. Association for Computational Linguistics.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erhard Schmidt. 1907. Zur theorie der linearen und nichtlinearen integralgleichungen. i. teil: Entwicklung willkürlicher funktionen nach systemen vorgeschriebener. *Mathematische Annalen*, 63:433–476.

M.A. Turk and A.P. Pentland. 1991. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Binxu Wang and Carlos R Ponce. 2021. A geometric analysis of deep generative image models and its applications. In *International Conference on Learning Representations*.

Xiaogang Wang and Xiaoou Tang. 2004. A unified framework for subspace face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1222–1228.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Runxin Xu, Fuli Luo, Baobao Chang, Songfang Huang, and Fei Huang. 2022. S^4-tuning: A simple cross-lingual sub-network tuning method. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 530–537, Dublin, Ireland. Association for Computational Linguistics.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5825–5832, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

Jiapeng Zhu, Ruili Feng, Yujun Shen, Deli Zhao, Zheng-Jun Zha, Jingren Zhou, and Qifeng Chen. 2021. Low-rank subspaces in gans. In *Advances in Neural Information Processing Systems*, volume 34, pages 16648–16658. Curran Associates, Inc.

A Theoretical Justification

In this section, we present Theorem 1 and the corresponding proof. We follow the same proving procedure in Piratla et al. (2020).

Theorem 1. *For any matrix $M \in \mathbb{R}^{d \times L}$, Algorithm 1 returns $\mu \in \mathbb{R}^d, M_s \in \mathbb{R}^{d \times r}, \Gamma \in \mathbb{R}^{L \times r}$ that minimize Equation 1 where $\mu \perp \mathrm{Span}(M_s)$.*

Proof. Algorithm 1 first obtains the best approximation of M with rank $r+1$ and $\mathbb{1}$ in its row space (Line 1-3). The orthogonal constraint $\mu \perp \mathrm{Span}(M_s)$ is then forced without obeying the low-rank property (Line 4-5).

To begin with, note that the optimization problem in Equation 1 is equivalent to the following:

$$\min_{\widehat{M}} \quad \left\| M - \widehat{M} \right\|_F^2$$
$$\text{s.t.} \quad \mathrm{rank}\left(\widehat{M}\right) \leq r + 1 \text{ and} \qquad (2)$$
$$\mathbb{1} \in \mathrm{Span}\left(\widehat{M}^\top\right).$$

Let $U, \Sigma, V = \mathrm{SVD}\left(M - \mu'\mathbb{1}^\top\right)$. We have that $\mathbb{1} \perp \mathrm{Span}\left(V^\top\right)$ given $\left(M - \mu'\mathbb{1}^\top\right)\mathbb{1} = 0$. Denote by $U_r\Sigma_r V_r^\top$ the top-r component of $U\Sigma V^\top$, by $\sigma_i(A)$ the i-th largest singular value of A and by A_i the best rank-i approximation of A.

The first step is to show that $\mu'\mathbb{1}^\top + U_r\Sigma_r V_r^\top$ minimizes the objective in Equation 2. Following the proof of Eckart-Young-Mirsky theorem for low-rank approximation (Schmidt, 1907; Eckart and Young, 1936), let $\widetilde{M} := M - \widehat{M}$ with any feasible $\widehat{M}$ fixed. We have

$$\sigma_i\left(\widetilde{M}\right) = \left\| \widetilde{M} - \widetilde{M}_{i-1} \right\|_F$$
$$= \left\| \widetilde{M} - \widetilde{M}_{i-1} \right\|_F + \left\| \widehat{M} - \widehat{M} \right\|_F$$
$$\geq \left\| \widetilde{M} + \widehat{M} - \widetilde{M}_{i-1} - \widehat{M} \right\|_F$$
$$= \left\| M - \widetilde{M}_{i-1} - \widehat{M} \right\|_F$$
$$\geq \min_{\bar{M}} \left\| M - \bar{M} \right\|_F,$$

where the minimum is taken over all $\bar{M}$ with $\mathrm{rank}(\bar{M}) = i + r$ and $\mathbb{1} \in \mathrm{Span}\left(\bar{M}^\top\right)$. By taking $\bar{M} = \mu'\mathbb{1}^\top + U_{i+r-1}\Sigma_{i+r-1}V_{i+r-1}^\top$, we have $\sigma_i\left(\widetilde{M}\right) \geq \sigma_{i+r}\left(U\Sigma V^\top\right)$ and therefore $\left\| M - \widehat{M} \right\|_F^2 \geq \left\| M - \mu'\mathbb{1}^\top - U_r\Sigma_r V_r^\top \right\|_F^2$.

Next, we find μ and M_s that meet the orthogonality constraint while preserving the low-rank structure. Suppose $\mu\mathbb{1}^\top + M_s\Gamma^\top = \mu'\mathbb{1}^\top + M_s'\Gamma'^\top$ with $\mu \perp \mathrm{Span}(M_s)$, we have that $\mu^\top\left(\mu\mathbb{1}^\top + M_s\Gamma^\top\right) = \|\mu\|^2 \mathbb{1}^\top$ which yields $\mu^\top = \|\mu\|^2 \left(\mu'\mathbb{1}^\top + M_s'\Gamma'^\top\right)^+ \mathbb{1}^\top$. $\qquad \square$

B Base Models

We evaluate the alignment methods based on a number of established pretrained multilingual models. We mainly build on the Transformers library (Wolf et al., 2020) for our experiments.

mBERT[5] Multilingual BERT (Devlin et al., 2019) is a transformer model (Vaswani et al., 2017) pretrained on Wikipedia, with the objective of Masked Language Modeling (MLM) and a shared vocabulary across all languages.

XLM[6] XLM (Conneau and Lample, 2019) also uses the MLM objective and the monolingual Wikipedia corpus for pretraining, with a larger model and a larger vocabulary.

XLM-R[7] XLM-R (Conneau et al., 2020a) follows a similar training procedure as XLM but collects the larger-scale CommonCrawl corpus.

LABSE[8] LABSE (Feng et al., 2022) is the state-of-the-art multilingual sentence encoder that leverages bilingual sentence pairs for pretraining.

Following previous works (Jawahar et al., 2019; Ruder et al., 2021) that observe certain intermediate layers of Transformer consistently outperform the last layer for cross-lingual tasks, we use the 8th layer for mBERT and XLM, and the 12th layer for XLM-R. We apply mean-pooling to obtain sentence embeddings as it is widely used (Conneau et al., 2020b; Muller et al., 2021). For LABSE as well as mBERT (X-X) and mBERT (En-En) used in LAReQA, we evaluate the alignment methods on the original sentence embeddings.

C Supplementary Results

In this section, we provide supplementary experimental results.

C.1 Hyperparameter Selection

For the considered baselines, we do not conduct sophisticated hyperparameter search given that it is non-trivial for LIR. To provide fair comparison, for LIR and LSAR that both have one single hyperparameter (the number of top principal components k and the number of basis vectors to span the low-rank subspace r), we exhaustively enumerate all
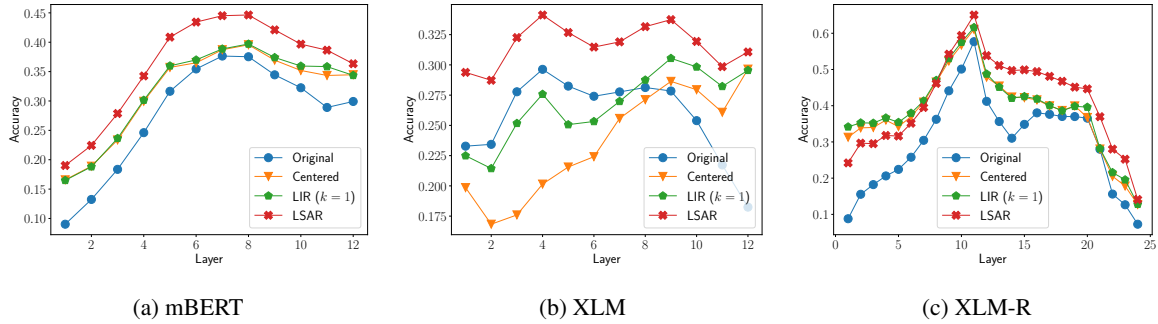
---

[5] https://huggingface.co/bert-base-multilingual-cased.
[6] https://huggingface.co/xlm-mlm-100-1280.
[7] https://huggingface.co/xlm-roberta-large.
[8] https://huggingface.co/sentence-transformers/LaBSE.

	(a) mBERT	(b) XLM	(c) XLM-R

Figure 6: Retrieval accuracy on Tatoeba (averaged over all 36 languages) at different layers.

	XQuAD-R		MLQA-R	
	En-En	X-X	En-En	X-X
Original	28.57	23.36	35.71	26.21
Centered	35.38	45.47	35.87	43.27
LIR ($k = 1$)	36.71	45.24	37.56	43.24
LIR ($k = 2$)	36.70	44.74	37.11	42.42
LIR ($k = 3$)	36.82	44.54	36.87	42.28
LSAR ($r = 1$)	30.51	26.38	36.79	28.79
LSAR ($r = 2$)	32.31	29.22	38.05	31.70
LSAR ($r = 3$)	34.05	31.99	39.00	35.28
LSAR	**40.95**	**46.39**	**40.70**	**44.02**

Table 6: Answer retrieval mAP (%) on XQuAD-R and MLQA-R of LAReQA (averaged over all languages), using Wiki-40B as the text resource.

	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
LIR ($k = 2$)	36.83	43.58	37.60	41.63
LIR ($k = 3$)	36.21	43.15	36.89	41.03
LSAR ($r = 1$)	30.50	26.27	36.68	28.59
LSAR ($r = 2$)	32.36	28.69	37.94	31.15
LSAR ($r = 3$)	34.20	31.49	38.82	34.46
LSAR	**41.13**	**45.89**	**40.55**	**43.32**

Table 7: Answer retrieval mAP (%) on XQuAD-R and MLQA-R of LAReQA (averaged over all languages), using OSCAR as the text resource.

values within a scope and report the best performances on the test data. Figure 7 shows the trend of accuracy on Tatoeba as the hyparameters change.

C.2 Wiki40-B Results

In this section we list the results of LAReQA (Table 6) and Amazon Reviews (Table 8-11) with Wiki-40B (Guo et al., 2020)[9] as the text resource. For Amazon Reviews, we also report the performances obtained in the last layers to reproduce those in Yang et al. (2021).

For Amazon Reviews, we determine the L2 regularization strength using a hyperparameter sweep on the 5-fold cross-validation routine, over the range between 1e-4 and 1e4 with 10 logarithmically spaced steps. This training procedure is implemented using the Scikit-Learn library (Buitinck et al., 2013).

C.3 OSCAR Results

The detailed results with OSCAR is provided in this section.

Tatoeba We report the results for all languages on Tatoeba in Table 17-20. Additionally, the complete set of results for clustering performance is shown in Table 12.

LAReQA We report the detailed results on LAReQA in Table 7. We omit listing all languages due to limited space.

Amazon Reviews We provide the results for all languages on Amazon Reviews in Table 13-16.

---

[9]https://www.tensorflow.org/datasets/catalog/wiki40b.

	Layer 8					Layer 12				
	en	de	fr	jp	avg.	em	de	fr	jp	avg.
Original	81.13	72.82	76.02	68.98	74.74	80.07	70.05	73.75	64.86	72.18
LIR ($k = 1$)	81.12	72.33	76.80	72.25	75.62	80.03	70.00	71.73	67.51	72.32
LIR ($k = 2$)	81.05	71.90	76.80	72.35	75.52	79.98	71.15	72.50	69.04	73.17
LIR ($k = 3$)	81.10	72.23	76.22	71.06	75.15	80.03	70.85	73.67	69.36	73.48
LSAR ($r = 1$)	81.12	72.77	75.87	72.30	75.51	79.98	71.17	73.68	71.15	73.99
LSAR ($r = 2$)	81.13	72.50	76.85	72.33	75.70	80.08	71.23	73.45	70.91	73.92
LSAR	81.12	72.43	76.67	72.36	75.64	79.87	70.10	71.95	68.69	72.65

Table 8: Classification accuracy (%) on Amazon Reviews (mBERT), using Wiki-40B as the text resource.

	Layer 8					Layer 12				
	en	de	fr	jp	avg.	em	de	fr	jp	avg.
Original	85.45	69.07	81.50	65.21	75.31	84.43	55.42	72.87	58.23	67.74
LIR ($k = 1$)	85.52	73.68	81.52	65.66	76.59	84.50	75.77	79.88	60.98	75.28
LIR ($k = 2$)	85.52	73.32	81.45	64.31	76.15	84.65	75.58	79.73	60.79	75.19
LIR ($k = 3$)	85.60	72.10	81.62	62.46	75.44	84.52	75.52	79.40	63.03	75.62
LSAR ($r = 1$)	85.53	70.98	81.52	66.44	76.12	84.45	56.75	75.20	66.64	70.76
LSAR ($r = 2$)	85.48	73.77	81.65	65.43	76.58	84.48	60.35	71.25	66.54	70.66
LSAR	85.50	73.78	81.63	65.41	76.58	84.48	75.78	79.57	64.99	76.21

Table 9: Classification accuracy (%) on Amazon Reviews (XLM), using Wiki-40B as the text resource.

	Layer 11					Layer 24				
	en	de	fr	jp	avg.	em	de	fr	jp	avg.
Original	84.33	78.32	82.30	76.35	80.32	90.55	78.08	83.57	67.14	79.84
LIR ($k = 1$)	84.33	82.47	81.68	80.18	82.17	90.53	88.67	89.88	86.16	88.81
LIR ($k = 2$)	84.45	82.18	82.10	80.08	82.20	90.62	88.48	88.27	85.61	88.25
LIR ($k = 3$)	84.33	81.40	83.08	78.48	81.82	90.67	88.55	88.40	85.61	88.31
LSAR ($r = 1$)	84.35	77.95	81.93	79.78	81.00	90.62	69.20	90.00	83.98	83.45
LSAR ($r = 2$)	84.33	82.52	81.17	80.53	82.14	90.60	88.73	79.18	79.30	84.45
LSAR	84.30	82.67	81.80	80.56	82.33	90.58	88.42	89.33	85.95	88.57

Table 10: Classification accuracy (%) on Amazon Reviews (XLM-R), using Wiki-40B as the text resource.

	en	de	fr	jp	avg.
Original	83.32	81.37	84.27	79.26	82.05
LIR ($k = 1$)	83.18	81.70	84.32	79.51	82.18
LIR ($k = 2$)	83.20	81.92	84.18	79.33	82.16
LIR ($k = 3$)	83.18	81.83	84.32	79.45	82.19
LSAR ($r = 1$)	83.32	81.30	84.28	79.21	82.03
LSAR ($r = 2$)	83.10	81.63	83.90	79.61	82.06
LSAR	83.27	81.77	83.95	79.75	82.18

Table 11: Classification accuracy (%) on Amazon Reviews (LABSE), using Wiki-40B as the text resource.

	mBERT	XLM	XLM-R	LABSE
Original	0.2815	0.5422	0.2457	0.0344
Centered	0.0975	0.2483	0.2004	0.0388
LIR ($k = 1$)	0.0900	0.1875	0.2203	0.0352
LSAR	0.0801	0.1320	0.0856	0.0306

Table 12: Clustering performance (NMI) of embeddings obtained by mBERT on Tatoeba.
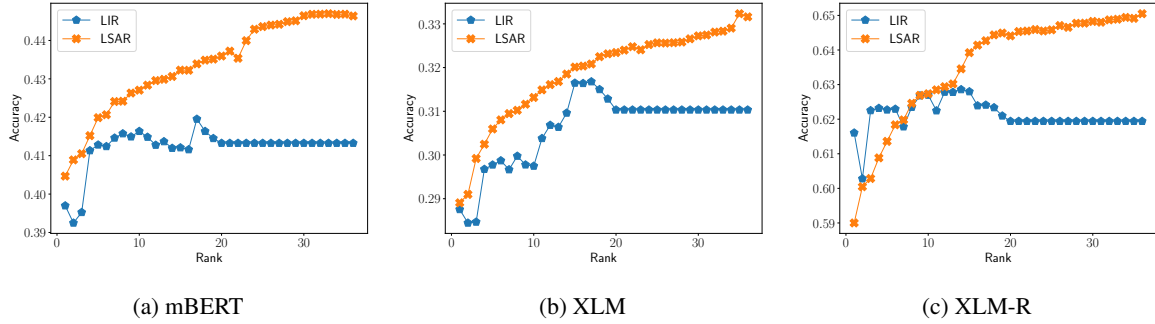
(a) mBERT   (b) XLM   (c) XLM-R

Figure 7: Retrieval accuracy on Tatoeba (averaged over all 36 languages) with different hyperparameters (k for LIR and r for LSAR). We observe that removing more principal components within each language for LIR does not result in better performances and can instead lead to information loss. For mBERT and XLM, the best k is found 17, whereas it is 14 for XLM-R. LSAR, however, consistently achieves the best results with $r = 36$ as larger subspaces encode more language-specific signals.

	Layer 8					Layer 12				
	en	de	fr	jp	avg.	em	de	fr	jp	avg.
Original	81.13	72.82	76.02	68.98	74.74	80.07	70.05	73.75	64.86	72.18
LIR ($k = 1$)	81.12	72.90	75.08	72.43	75.38	80.07	71.08	71.40	67.11	72.42
LIR ($k = 2$)	81.03	71.47	70.58	66.09	72.29	79.97	69.35	72.07	66.29	71.92
LIR ($k = 3$)	80.85	68.67	74.38	67.53	72.86	79.88	66.10	69.80	66.59	70.59
LSAR ($r = 1$)	81.25	72.78	75.80	72.48	75.58	79.98	71.03	73.62	70.45	73.77
LSAR ($r = 2$)	81.27	72.57	75.85	72.30	75.49	80.07	71.12	73.48	70.11	73.69
LSAR	81.15	72.90	75.22	71.68	75.24	79.88	70.80	71.70	67.79	72.54

Table 13: Classification accuracy (%) on Amazon Reviews (mBERT), using OSCAR as the text resource.

	Layer 8					Layer 12				
	en	de	fr	jp	avg.	em	de	fr	jp	avg.
Original	85.45	69.07	81.50	65.21	75.31	84.43	55.42	72.87	58.23	67.74
LIR ($k = 1$)	85.58	77.57	80.05	59.74	75.74	84.52	75.75	80.20	55.26	73.93
LIR ($k = 2$)	85.40	76.72	79.82	60.86	75.70	84.48	75.57	77.95	55.46	73.36
LIR ($k = 3$)	85.15	77.42	81.07	51.51	73.79	84.48	74.55	76.13	51.26	71.61
LSAR ($r = 1$)	85.47	69.08	81.42	63.78	74.94	84.50	56.33	74.63	66.84	70.58
LSAR ($r = 2$)	85.37	74.53	81.60	61.88	75.84	84.50	57.75	72.80	66.86	70.48
LSAR	85.45	77.15	80.25	58.24	75.27	84.62	75.87	80.65	57.14	74.57

Table 14: Classification accuracy (%) on Amazon Reviews (XLM), using OSCAR as the text resource.

	Layer 11					Layer 24				
	en	de	fr	jp	avg.	em	de	fr	jp	avg.
Original	84.33	78.32	82.30	76.35	80.32	90.55	78.08	83.57	67.14	79.84
LIR ($k = 1$)	84.32	82.55	77.82	79.93	81.15	90.53	88.85	87.67	86.11	88.29
LIR ($k = 2$)	84.42	82.27	78.15	79.45	81.07	90.63	89.12	85.93	85.86	87.89
LIR ($k = 3$)	84.33	81.05	77.57	79.16	80.53	90.68	89.85	84.68	86.30	87.88
LSAR ($r = 1$)	84.32	78.80	82.12	80.66	81.47	90.55	83.47	77.67	80.86	83.14
LSAR ($r = 2$)	84.32	82.55	82.08	80.53	82.37	90.55	87.63	76.57	77.66	83.10
LSAR	84.27	82.60	77.85	80.28	81.25	90.57	89.37	88.03	86.01	88.50

Table 15: Classification accuracy (%) on Amazon Reviews (XLM-R), using OSCAR as the text resource.

	en	de	fr	jp	avg.
Original	83.32	81.37	84.27	79.26	82.05
LIR ($k = 1$)	83.40	81.85	82.62	79.81	81.92
LIR ($k = 2$)	83.28	80.92	78.37	78.73	80.32
LIR ($k = 3$)	82.88	78.92	78.82	78.85	79.87
LSAR ($r = 1$)	83.07	81.52	83.88	79.20	81.92
LSAR ($r = 2$)	83.02	82.10	83.55	79.66	82.08
LSAR	83.13	81.92	83.18	79.48	81.93

Table 16: Classification accuracy (%) on Amazon Reviews (LABSE), using OSCAR as the text resource.

	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr
Original	38.90	24.50	48.80	17.00	75.40	29.80	64.10	28.10	25.50	41.20	39.00	64.30
Centered	40.90	27.30	48.50	17.30	74.70	35.10	66.40	29.60	27.40	43.70	40.30	65.30
LIR ($k = 1$)	41.00	27.20	48.60	17.90	74.90	35.10	66.40	30.10	27.70	44.00	40.50	64.90
LSAR	44.70	31.80	55.00	21.90	79.00	38.70	71.20	35.30	32.00	49.80	46.40	69.10
	he	hi	hu	id	it	ja	jv	ka	kk	ko	ml	mr
Original	40.10	34.80	36.90	53.50	57.30	40.90	17.56	19.57	27.13	36.00	17.90	20.10
Centered	41.50	35.40	41.40	53.40	58.30	41.60	18.54	23.32	30.96	38.70	27.66	23.00
LIR ($k = 1$)	41.70	35.40	41.60	53.70	58.20	41.90	18.05	23.73	30.96	38.80	28.82	23.00
LSAR	45.70	43.90	46.00	60.00	61.90	51.00	24.88	28.28	34.09	45.30	36.83	26.40
	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh
Original	63.70	68.40	59.40	10.77	13.36	14.10	13.69	16.00	32.90	30.80	61.00	68.60
Centered	64.30	69.50	62.40	12.56	14.33	14.96	17.15	18.10	38.20	31.40	62.20	69.00
LIR ($k = 1$)	65.10	69.30	62.10	12.31	14.33	14.96	17.15	18.20	38.20	32.10	62.00	69.20
LSAR	69.20	73.10	67.20	14.36	18.57	21.37	21.72	22.00	41.90	38.00	67.10	73.30

Table 17: Retrieval accuracy (%) on Tatoeba for each language (mBERT), using OSCAR as the text resource.

	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr
Original	34.20	17.80	34.80	5.70	62.20	24.90	56.00	18.40	11.90	30.50	28.10	52.80
Centered	30.30	17.30	35.30	5.00	62.20	22.50	53.50	19.20	14.70	29.90	31.30	49.20
LIR ($k = 1$)	32.20	18.20	37.30	5.80	65.10	25.60	54.10	21.10	16.60	31.00	32.00	51.70
LSAR	37.50	20.10	42.40	9.90	68.20	30.50	58.80	25.50	22.00	35.00	36.10	55.10
	he	hi	hu	id	it	ja	jv	ka	kk	ko	ml	mr
Original	31.20	15.70	29.50	44.60	52.20	32.20	19.51	22.12	14.26	25.20	0.58	6.30
Centered	30.00	14.50	30.00	45.10	49.90	28.60	17.56	19.71	14.78	22.70	0.44	5.50
LIR ($k = 1$)	31.40	17.40	31.20	45.40	50.60	31.90	19.02	21.85	16.70	24.50	0.87	6.20
LSAR	34.10	24.30	36.70	49.20	55.10	36.80	22.44	24.80	20.87	29.30	4.95	10.70
	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh
Original	55.00	58.40	44.20	8.97	1.63	5.56	27.74	12.40	24.90	17.80	45.70	39.70
Centered	55.60	58.10	42.50	6.92	2.28	5.13	18.43	14.60	27.70	16.40	43.70	36.00
LIR ($k = 1$)	57.30	58.80	43.60	9.49	2.28	5.56	23.91	15.20	28.80	17.20	45.20	40.10
LSAR	59.70	61.90	47.60	11.79	6.84	11.54	32.66	20.10	33.50	22.90	52.00	42.90

Table 18: Retrieval accuracy (%) on Tatoeba for each language (XLM), using OSCAR as the text resource.

	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr
Original	58.20	47.50	71.60	43.00	88.80	61.80	75.70	52.20	35.80	70.50	71.60	73.70
Centered	59.30	49.60	75.00	45.30	90.90	65.80	76.60	57.10	45.80	72.10	78.40	73.00
LIR ($k = 1$)	59.80	50.30	75.30	46.10	90.70	66.30	77.20	57.50	47.00	72.60	78.80	73.80
LSAR	65.20	55.00	76.50	52.60	91.60	71.30	80.90	60.90	52.00	75.90	78.90	77.50
	he	hi	hu	id	it	ja	jv	ka	kk	ko	ml	mr
Original	66.40	72.20	65.40	77.00	68.30	60.60	14.15	52.28	48.52	61.40	65.36	56.80
Centered	69.10	74.10	67.90	80.00	70.60	62.50	21.95	62.60	49.57	63.00	70.01	60.30
LIR ($k = 1$)	69.50	75.00	68.20	80.40	71.50	62.60	20.98	63.00	50.78	63.50	69.87	61.20
LSAR	71.80	79.40	72.70	81.50	73.70	68.20	26.34	61.53	55.65	69.70	76.71	67.60
	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh
Original	80.80	82.20	74.10	20.26	26.38	35.90	29.38	36.70	65.70	23.40	74.70	68.30
Centered	81.80	81.50	78.20	24.10	30.62	41.45	30.29	37.30	74.00	26.90	79.70	72.60
LIR ($k = 1$)	82.10	82.20	78.80	25.64	31.60	41.88	31.02	37.60	74.50	27.00	80.40	73.10
LSAR	84.10	84.30	79.00	26.92	36.16	44.02	35.04	47.00	75.50	32.90	79.90	73.80

Table 19: Retrieval accuracy (%) on Tatoeba for each language (XLM-R), using OSCAR as the text resource.

	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr
Original	97.70	90.60	95.50	91.60	99.30	96.70	98.10	98.00	95.40	96.30	97.00	96.10
Centered	97.60	90.40	95.60	91.60	99.30	96.60	98.30	98.10	95.70	96.20	97.20	96.30
LIR ($k = 1$)	97.70	90.40	95.60	91.60	99.30	96.80	98.10	98.10	95.80	96.10	97.00	96.30
LSAR	97.40	90.90	95.40	91.60	99.30	96.60	98.20	97.90	95.60	95.90	97.10	96.30
	he	hi	hu	id	it	ja	jv	ka	kk	ko	ml	mr
Original	92.40	97.90	97.00	95.60	95.30	96.40	85.37	95.71	91.13	94.10	98.98	95.00
Centered	92.10	97.90	97.10	95.80	95.20	96.70	87.80	95.58	91.30	94.20	99.13	95.00
LIR ($k = 1$)	92.10	97.90	97.00	95.60	95.40	96.50	87.80	95.71	91.83	94.00	99.13	95.20
LSAR	92.40	97.80	97.10	95.80	95.40	96.50	85.85	95.71	91.65	93.90	99.13	94.80
	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh
Original	97.50	95.70	95.30	89.49	90.23	98.29	97.08	98.00	98.20	96.00	97.80	96.10
Centered	97.70	95.60	95.00	89.23	90.23	98.72	97.26	97.90	98.20	95.70	97.90	96.00
LIR ($k = 1$)	97.70	95.70	95.20	90.26	90.55	98.72	97.45	98.00	98.30	95.90	97.80	96.00
LSAR	97.50	96.00	95.40	90.26	90.55	98.72	96.90	97.80	98.30	96.00	97.70	96.20

Table 20: Retrieval accuracy (%) on Tatoeba for each language (LABSE), using OSCAR as the text resource.