CONSTRAINED DENSITY MATCHING AND MODELING FOR EFFECTIVE CONTEXTUALIZED ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual representations pre-trained with monolingual data offer task performances considerably unequal between languages. While this has been tackled through the lens of contextualized alignments, these techniques require large parallel data, thereby leaving under-represented language communities behind. In this work, we analyze the limitations according to which previous alignments become very resource-intensive, *viz.,* (i) the inability to sufficiently leverage data and (ii) that alignments are not trained properly. To address them, we present density based approaches to perform alignments, and we complement them with our validation criteria accounting for both intrinsic and extrinsic task performances. Our experiments encompass 16 alignments (including ours) evaluated across 6 language pairs, synthetic and 4 NLP tasks. We demonstrate that our solutions are particularly effective in the scenarios of limited and no parallel data. More importantly, we show, both theoretically and empirically, the advantages of our boostrapping procedures, by which unsupervised approaches rival supervised counterparts.

1 INTRODUCTION

Multilingual text encoders such as m-BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have been profiled as the de facto solutions to modeling languages at scale. However, such encoders pre-trained with monolingual data often produce very different representations for parallel data, particularly in low-resource and distant languages settings (Zhao et al., 2020a) as the induced multilingual subspaces are mis-aligned. As remedies, supervised alignment techniques have received large attention. Researchers use them to rectify multilingual representations post-hoc with cross-lingual supervision (Aldarmaki & Diab, 2019; Cao et al., 2020; Zhao et al., 2020b; Chi et al., 2021), but their utility is limited to high-resource languages with large parallel data. While unsupervised techniques remove the need of parallel data and thus promise unlimited applicability for all language pairs, they have surprisingly been little studied in contextualized, multilingual representations.

In this work, we particularly address the scenarios of limited and no parallel data. In **supervised** settings, we tackle two limitations leading to the ineffectiveness of previous alignments, namely, (i) the inability to sufficiently leverage data, i.e., to model data density with limited parallel data and (ii) that alignments are not properly trained due to a lack of validation criteria[1]. To address them, we model density based alignments as follows: First, we frame the problem of aligning multilingual representations as density matching. Next, we use Normalizing Flows (Dinh et al., 2017) to perform density estimation, by which we leverage data (be it parallel or not). Finally, we introduce two validation criteria accounting for task performances, to prevent models from overfitting.

In **unsupervised settings**, we reuse the above paradigm that combines density based approaches and validation criteria. However, we identify a statistical issue of density matching in the unsupervised case, i.e., that it only leads to a weak notion of equality of multilingual subspaces, *viz.*, equality in distribution. To address this, we introduce boostrapping procedures promoting equality of multilingual subspaces, from which we derive cross-lingual supervision from non-parallel data (see §2).

We thoroughly evaluate 16 alignment techniques (including ours) across 6 language pairs, synthetic and 4 NLP tasks. Our major findings are summarized as follows:

[1]Alignments such as those of Wu & Dredze (2020) and Cao et al. (2020) have been trained for several epochs without the use of any criteria for model selection; this comes at the risk of mistraining, given that task performances are not taken into account.

- Combining density matching and modeling leads to the best alignment design in both supervised and unsupervised settings. Unsupervised alignments integrated in boostrapping procedures rival supervised counterparts. However, as exposed in simulation setups, we find that such alignments have trouble in generalization to unseen words.

- Our validation criteria are better alternatives to intrinsic data for model selection, as the latter's results do not correlate with extrinsic task results. Further, we find intra-task results correlate negatively in about 30% of setups, showing that validation criteria should address a performance trade-off between tasks by taking all these tasks performances into account.

2 Related Work

Recent advances in multilingual representations (e.g., m-BERT and XLM-R) boost performances of cross-lingual systems in NLP; however, such systems exhibit weak(er) performances for distant languages (Pires et al., 2019; Wu & Dredze, 2019) and for languages with resource disparity (Lauscher et al., 2020; Zhao et al., 2020a). To address this, contextualized alignments have been proposed in the literature: Aldarmaki & Diab (2019) show that language-independent rotation can linearly rectify m-BERT representations, while Cao et al. (2020) find that jointly aligning multiple languages performs better. Zhao et al. (2020b) show that removing language bias in multilingual representations mitigates misalignments between languages. Recently, Mengzhou et al. (2021) suggest to use gradient based alignments, showing that they are effective for unseen languages in XLM-R. Alqahtani et al. (2021) use optimal transport to finetune multilingual representations, while other works (Hu et al., 2021; Chi et al., 2021) finetune them with translation language modeling. However, these efforts have predominantly focused on supervised, resource-intensive alignments that use 250k-2M parallel sentences per language pair for considerable improvements. As early attempts to remove the use of parallel data, Libovický et al. (2020) and Zhao et al. (2020a) find that applying vector space normalizations alone is helpful to yield language-neural representations.

However, there still lacks a systematic study on unsupervised alignments for contextualized multilingual representations, while these have been long researched in the context of static embeddings. The latter often uses iterative procedures to derive bilingual lexicons (as cross-lingual supervision) from monolingual data in two steps: (i) inducing seed dictionaries using different approaches such as adversarial learning (Lample et al., 2018), similarity based heuristics (Artetxe et al., 2018b) and identical strings (Smith et al., 2017; Artetxe et al., 2017) and (ii) applying Procrustes to augment induced lexicons (Artetxe et al., 2018b; Lample et al., 2018) in an iterative fashion. In this work, we present a principled, iterative procedure for unsupervised, contextualized alignments: First, we use density based approaches to induce bilingual lexicons. Next, we apply our boostrapping procedures that theoretically lead to equality of multilingual subspaces, to iteratively augment lexicons. Finally, we complement the first two steps with our validation criteria accounting for all tasks that we investigate.

3 Contextualized Alignment

Let two random variables X and Y with densities P_X and P_Y describe respective contextual word embeddings in two languages ℓ_1 and ℓ_2, with Ω_{ℓ_1} and Ω_{ℓ_2} as two lexicons. Each occurrence of a word is associated to a separate entry in the lexicons. X maps all entries in Ω_{ℓ_1} to real-valued m-dimensional embedding vectors, denoted by $X : \Omega_{\ell_1} \to \mathbb{R}^m$, and similarly for Y. A bilingual lexicon Ω describes a set of translations between Ω_{ℓ_1} and Ω_{ℓ_2}.

Empirical inference. Assume a function $f : \mathbb{R}^m \to \mathbb{R}^m$ perfectly maps m-dimensional embedding vectors from X to Y. As standard in machine learning, a mapping function f_θ can be empirically inferred from data, with θ as model parameters. To do so, we assume data $\mathbf{M}_{\ell_1} \in \mathbb{R}^{n \times m}$ and $\mathbf{M}_{\ell_2} \in \mathbb{R}^{n \times m}$ are given, which are two sets of contextual word embeddings with a common size of n for simplicity. Let a permutation matrix $\mathbf{P} \in \{0,1\}^{n \times n}$ ($\mathbf{P}\mathbf{1}_n = \mathbf{1}_n$ and $\mathbf{P}^\top \mathbf{1}_n = \mathbf{1}_n$) be a realization of Ω, serving as cross-lingual supervision when available. A random variable $\tilde{Y}$ with the density $P_{\tilde{Y}}$ is a prediction of Y given X, namely $\tilde{Y} = f_{X \to Y}(X)$ (we use subscripts to denote the mapping direction).

3.1 SUPERVISED ALIGNMENT

When parallel data is available, one can induce a permutation matrix $\mathbf{P}$ from parallel data either by bilingual experts or by word alignment tools such as FastAlign (Dyer et al., 2013) or SimAlign (Jalili Sabet et al., 2020). Here, we introduce a density based mapping function. As an alternative, we frame learning objectives of contextualized alignment as density matching between $P_{\tilde{Y}}$ and P_Y:

$$\begin{aligned}\mathrm{KL}(P_{\tilde{Y}}, P_Y) &= \mathrm{CE}(P_{\tilde{Y}}, P_Y) - \mathbb{E}_{y\sim P_Y}[\log P_Y(y)] \\ &= \|f_{X\to Y}(\mathbf{M}_{\ell_1}) - \mathbf{P}\mathbf{M}_{\ell_2}\|^2 - \mathbb{E}_{y\sim P_Y}[\log P_X(f^{-1}_{X\to Y}(y))|\det(\nabla_\theta f^{-1}_{X\to Y}(y))|] \end{aligned} \tag{1}$$

where $f_{X\to Y}$ is the trainable mapping function from X to Y. As P_Y is intractable to compute, previous work in supervised alignments always minimizes the cross-entropy term alone by solving the least squares problem. Note that the density P_Y can be rewritten to $P_X(f^{-1}_{X\to\tilde{Y}}(y))|\det(\nabla_\theta f^{-1}_{X\to\tilde{Y}}(y))|$ by using the change-of-variable rule (assuming f is an invertible function), but the density $P_{\tilde{Y}}$ is still intractable to compute given the unknown density P_X. To address this issue, we introduce a generative model, namely Real-NVP (Dinh et al., 2017) as use case of Normalizing Flows (Dinh et al., 2015; Rezende & Mohamed, 2015), which learns a mapping between data points with an unknown distribution and with the normalized Gaussian distribution through the change-of-variable rule. From this, one can infer data likelihood of an unknown complex distribution using a simple Gaussian distribution. This technique has been shown useful in many cross-domain applications, e.g., image-to-image translation (Grover et al., 2020; Gong et al., 2019) and machine translation (Setiawan et al., 2020).

In particular, we introduce a latent variable Z with the normalized density P_Z, namely $\mathcal{N}(0,\mathbf{I})$, and we let Z share a common vector space with X and Y. We use Real-NVP to infer the likelihood of data from Y using P_Z, denoted by $P_Y(y) = P_Z(f^{-1}_{Z\to Y}(y))|\det(\nabla_\theta f^{-1}_{Z\to Y}(y))|$ with $f_{Z\to Y}$ as a trainable mapping function from Z to Y. Finally, we rewrite the entropy term in Eq. 1 to:

$$\mathbb{E}_{y\sim P_Y}[\log P_Y(y)] = \mathbb{E}_{y\sim P_Y}[\log \mathcal{N}(f^{-1}_{Z\to Y}(y), 0, \mathbf{I})] + \mathbb{E}_{y\sim P_Y}[\log|\det(\nabla_\theta f^{-1}_{Z\to Y}(y))|] \tag{2}$$

In this work, we perform a dual form of density matching based on JS divergence, from which we take into account density estimation (modeling) for both P_X and P_Y. We omit the definition for simplicity.

3.2 UNSUPERVISED ALIGNMENT

When $\mathbf{P}$ is not given due to a lack of parallel data, we apply adversarial learning to align two densities $P_{\tilde{Y}}$ and P_Y. As standard in adversarial training, we involve a min-max game between two components to perform density matching: (a) a discriminator that distinguishes source and target word embeddings after mapping them and (b) a mapping function that aligns source and target word embeddings and fools the discriminator. Here, we use a popular adversarial approach, the Wasserstein GAN (Arjovsky et al., 2017). It minimizes the Earth Mover distance (EMD) between $\tilde{Y}$ and Y to align their densities $P_{\tilde{Y}}$ and P_Y. To better leverage data, we include density estimation based on Real-NVP in the procedure of adversarial training. By doing so, we maximize likelihood of data from X and Y using estimated densities P_X and P_Y (see Eq. 2). Taken together, we denote our density based learning objectives by:

$$\mathrm{EMD}(P_{\tilde{Y}}, P_Y) = \min_{f_{X\to Y}} \max_{h_\phi} \mathbb{E}_{y\sim P_Y}[h_\phi(y)] - \mathbb{E}_{\tilde{y}\sim P_{\tilde{Y}}}[h_\phi(\tilde{y}))] + \mathbb{E}_{y\sim P_Y}[\log P_Y(y)] \tag{3}$$

where h_ϕ is a 1-Lipschitz constrained discriminator, $\tilde{y} = f_{X\to Y}(x)$ maps embeddings from X to Y. Note that $f_{X\to Y}$ is the composition of $f_{X\to Z}$ and $f_{Z\to Y}$, and the last entropy term is to maximize log-likelihood of data from Y. As in the supervised case, we use a dual form of Eq. 3.

Bootstrapping procedure. After adversarial learning Y and $\tilde{Y}$ are ideally equal in distribution, denoted by $Y \overset{\text{dist}}{=} \tilde{Y}$. However, this is not sufficient. For instance, let $Y \sim \mathrm{Uniform}(-1,1)$ and $\tilde{Y} = -Y$. Then, Y and Y are equal in distribution, but they are identical only at the origin. As remedies, we derive two conditions to further improve alignments, promoting the equality of Y and $\tilde{Y}$.

Proposition 1. *Given $Y \overset{\text{dist}}{=} \tilde{Y}$, $\tilde{Y}$ and Y are equal if one of the following conditions is met:*

(i) $Y = \mathbf{U}\tilde{Y}$, *where* $\mathbf{U}$ *is invertible and* $\mathbf{U}_{ij} \geq 0\ \forall i, j$.

(ii) $\mathrm{cor}(\tilde{Y}_i, Y_i) = 1$ *for* $\forall i$, *where* $\tilde{Y}_i$ *and* Y_i *represent the i-th component in* $\tilde{Y}$ *and* Y.

Proof. (i) $P(Y \leq y) = P(\tilde{Y} \leq y)$ for all y, due to $Y \overset{\text{dist}}{=} \tilde{Y}$. If $Y = \mathbf{U}\tilde{Y}$, then $P(\tilde{Y} \leq y) = P(Y \leq y) = P(\mathbf{U}\tilde{Y} \leq y)$. If $\mathbf{U} \geq 0$, then $P(\mathbf{U}\tilde{Y} \leq y) = P(\tilde{Y} \leq \mathbf{U}^{-1}y)$. Thus, $P(\tilde{Y} \leq \mathbf{U}^{-1}y) = P(\tilde{Y} \leq y)$ for all y. This implies that $\mathbf{U} = \mathbf{I}$. Thus, $\tilde{Y} = Y$.

(ii) If $\text{cor}(\tilde{Y}_i, Y_i) = 1$ for $\forall i$, then $\text{Var}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})] = 0$, thus $\mathbb{E}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})^2] - \mathbb{E}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})]^2 = 0$. However, the second term equals to 0 by using $\mathbb{E}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})] = \frac{\mathbb{E}[\tilde{Y}_i]}{\sigma_{\tilde{y}_i}} - \frac{\mathbb{E}[Y_i]}{\sigma_{y_i}} = 0$ due to $\tilde{Y}_i \overset{\text{dist}}{=} Y_i$. Thus, $\mathbb{E}[(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})^2] = 0$, and this implies $\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} = \frac{Y_i}{\sigma_{y_i}}$ since the non-negative $(\frac{\tilde{Y}_i}{\sigma_{\tilde{y}_i}} - \frac{Y_i}{\sigma_{y_i}})^2$ must be zero if its expectation is 0. Note that $\sigma_{\tilde{y}_i} = \sigma_{y_i}$ due to $\tilde{Y}_i \overset{\text{dist}}{=} Y_i$. This implies that $\tilde{Y}_i = Y_i$ for $\forall i$. □

To realize the conditions in Prop. 1, we need additional notation and a lemma. Let $\mathbf{M}_X$, $\mathbf{M}_Y$ be embeddings from X and Y, and $\mathbf{M}_{\tilde{Y}}$=$f_\theta(\mathbf{M}_X)$.

Lemma 2. *If* $\mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^\intercal = \mathbf{M}_Y\mathbf{M}_Y^\intercal$ *and* $\mathbf{M}_{\tilde{Y}}$ *is invertible, then* $Y = \mathbf{U}\tilde{Y}$.

Proof. If $\mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^\intercal = \mathbf{M}_Y\mathbf{M}_Y^\intercal$ and $\mathbf{M}_{\tilde{Y}}$ is invertible, then $\mathbf{M}_Y = \mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^{-1}\mathbf{M}_Y$. Let $\mathbf{U} = \mathbf{M}_{\tilde{Y}}^{-1}\mathbf{M}_Y$. Then, $\mathbf{M}_Y = \mathbf{M}_{\tilde{Y}}\mathbf{U}$. If this holds for all $\mathbf{M}_{\tilde{Y}}$ and $\mathbf{M}_Y$, then $Y = \mathbf{U}\tilde{Y}$. □

In the following, we refer to our two realizations of Prop. 1 as constraints added to to Eq. 3, from which we promote the equality of Y and $\tilde{Y}$. Moreover, we discuss their connections with canonical correlation and with language isomorphism (see §A.1 (appendix)).

Graph structure. We describe $\mathbf{M}_{\tilde{Y}}$ and $\mathbf{M}_Y$ as m-dimensional vertices in two graphs, and denote their corresponding (weighted) adjacency matrices by $\mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^\intercal$ and $\mathbf{M}_Y\mathbf{M}_Y^\intercal$. Then, we minimize the difference between them. We use this to realize Prop.1(i)[2] according to Lemm. 2. The objective becomes:

$$\text{EMD}(P_{\tilde{Y}}, P_Y) + \|\mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^\intercal - \mathbf{M}_Y\mathbf{M}_Y^\intercal\|^2 \tag{4}$$

Cross-correlation. We maximize Pearson cross-correlation between (demeaned) $\mathbf{M}_{\tilde{Y}}$ and $\mathbf{M}_Y$, and use this to realize Prop.1(ii). The objective becomes:

$$\text{EMD}(P_{\tilde{Y}}, P_Y) + \|\frac{\text{diag}(\mathbf{M}_{\tilde{Y}}^\intercal\mathbf{M}_Y)}{\text{diag}(\mathbf{M}_{\tilde{Y}}^\intercal\mathbf{M}_{\tilde{Y}})\text{diag}(\mathbf{M}_Y^\intercal\mathbf{M}_Y)} - \vec{1}\|^2 \tag{5}$$

Regarding $\mathbf{M}_Y, \mathbf{M}_{\tilde{Y}}$, we use CSLS (Lample et al., 2018) to induce them, and take them as input arguments of our realizations to guide learning objectives. We refer to this as boostrapping procedures[3] (see Algorithm 1).

Algorithm 1 Bootstrapping Procedures

```
M_X, M_Y ← population word embeddings of X and Y
n ← number of boostrapping iterations
f_{X→Y} ← an identity matrix as initial mapping function
for each of n iterations do
    M_Ỹ ← f_{X→Y}(M_X)
    P ← CSLS(M_Y, M_Ỹ)                          ▷ induce permutation matrix
    f_{X→Y} ← EMD(P_Ỹ, P_Y) + g(M_Y, PM_Ỹ)      ▷ g is a realization (see Eq. 4+5)
end for
return f_{X→Y}
```

[2]Prop.1(i) is not strictly met, as it is not trivial to guarantee $\mathbf{U} \geq 0$, i.e., $\mathbf{M}_{\tilde{Y}}^{-1}\mathbf{M}_Y \geq 0$. This might explain why graph structure performs worse than cross-correlation in our experiments.

[3]Our initial study finds taking the difference between $\mathbf{M}_Y$ and $\mathbf{P}\mathbf{M}_{\tilde{Y}}$ as a realization leads to poor results, as it is very sensitive to noises in induced permutation matrices. We set n to 10 in simulation and to 3 on real data.

4 EXPERIMENTS

Unlike static embeddings, contextual embeddings have not yet been investigated in intrinsic tasks such as Bilingual Lexicon Induction (BLI). We address this by creating simulated data in §4.3, on which we contrast findings from static and contextual embeddings and expose limitations of alignment techniques (including ours). In §4.4, we compare the alignment techniques across 6 language pairs and 4 real cross-lingual NLP tasks.

4.1 BASELINES AND OURS

Supervised alignments. (a) Rotation (Aldarmaki & Diab, 2019; Zhao et al., 2020a): a linear orthogonal-constrained transformation; (b) GBDD (Zhao et al., 2020b): subtracting a global language bias vector from multilingual representations; (c) FCNN: an architecture that contains three fully-connected layers followed by a tanh activation function each; (d) a popular approach Joint-Align (Cao et al., 2020): jointly aligning multiple languages using fine-tuning; (e) recent InfoXLM (Chi et al., 2021): finetuning multilingual representations with translation language modeling and contrastive learning; and (f) our density based approach Real-NVP. Except for Joint-Align and InfoXLM, all other approaches have been trained separately for different language pairs.

Unsupervised alignments. (a) MUSE: an unsupervised variant of Rotation (Lample et al., 2018); (b) VecMap: a heuristic unsupervised approach that assumes word translations have similar distributions of word similarities (Artetxe et al., 2018a); (c) vector space normalization (Zhao et al., 2020a): removing language-specific means and variances of multilingual representations and (d) our GAN-Real-NVP: an unsupervised variant of Real-NVP. For boostrapping based unsupervised approaches, we denote them by [Method]+Cross-Correlation, [Method]+Graph-Structure and [Method]+Procrustes, where [Method] is MUSE or GAN-Real-NVP; Procrustes is a popular bootstrapping procedure commonly used for static embeddings. Both MUSE and VecMap are popular unsupervised alignments in static embeddings.

4.2 VALIDATION CRITERION

We examine two unsupervised validation criteria for model selection during training, and compare them with no-criteria (i.e., training for several epochs) in both supervised and unsupervised settings. Our criteria are measured based on the 30k most confident word translations induced by CSLS.

- *Semantic criterion* has been established in static embeddings. Lample et al. (2018) select the 10k most frequent source words and use them to generate target translations. Next, they average cosine similarities on word translations and use it as a validation criterion, and find it correlates strongly with results on validation sets.
- We measure the difference between two ordered lists of singular values obtained from source and target word embeddings. While Dubossarsky et al. (2020) use this to measure language isomorphism, we refer to it as a *structural criterion*.

4.3 SIMULATION

BLI is a popular intrinsic task used to evaluate static embeddings, as it covers over 100 language pairs and evaluates the word alignment itself other than its impact on extrinsic tasks. BLI evaluations induce bilingual lexicons according to similarities of word embeddings and compare them with gold lexicons. However, contextual embeddings lack such evaluations. As argued in Artetxe et al. (2020), when not evaluated under similar conditions, the lessons learned from static embeddings cannot transfer to contextual ones. To address this, we use data simulation to create synthetic data, a form of contextual extension of BLI (CBLI), to intrinsically evaluate contextual embeddings. We split data to train, validation and test sets, on which we compare induced bilingual lexicons with gold ones. Task performances are measured in Precision@K as in BLI evaluation. Our creation procedure is two-fold: First, we produce source embeddings sampled from a two-dimensional Gaussian mixture distribution. Next, we perform linear and non-linear transformations on them, to produce target embeddings, by which we mimic different typological language pairs (see example in Fig. 4 (appendix)).

When constructing simulation setups, we adjust three parameters: (a) the number of occurrences k for a word, $k \in \{5, 10, \ldots, 100\}$; we use 20 words in all setups, (b) the degree of language isomorphism

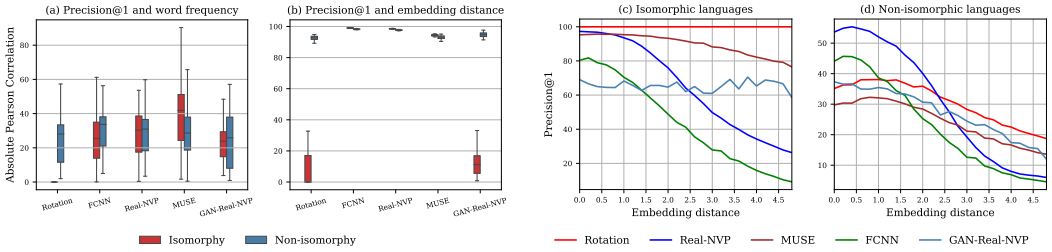

Figure 1: Pearson correlation between task performance and (a) word frequency (per word the number of occurrences) and (b) similarity between train and test domains (distances between embeddings in CBLI train and test sets). We set frequency bins $k \in \{5, 10, \ldots, 100\}$, and set similarity bins $\epsilon \in \{0, 0.2, \ldots, 5\}$. We set t to 1 in all isomorphic settings, and t to 5 as use case in non-isomorphic settings. (c)+(d) compares generalization abilities of approaches. Results are averaged across 10 runs.

$t \in \{1, \ldots, 10\}$—which mimics different language pairs and (c) the distance ϵ between embeddings in train and test sets, $\epsilon \in \{0, 0.2, \ldots, 5\}$—which reflects different similarities between train and test domains. For the i-th word and its occurrences, their embeddings are sampled from a Gaussian distribution $\mathcal{N}(\mu_i, \mathbf{I})$ for train sets and from a $\mathcal{N}(\mu_i + \epsilon, \mathbf{I})$ for validation and test sets. μ_i denotes a mean vector sampled uniformly from $[-5, 5]$ for each component. For isomorphic languages ($t = 1$), we transform source to target embeddings using an orthogonal matrix. For non-isomorphic languages ($t > 1$), we combine translation and rotation to perform non-linear transformations.

Generalization to unseen words. While Czarnowska et al. (2019) finds that word frequency has a big impact on task performance for static embeddings, Fig. 1 (a)+(b) show that, in the contextual case, task performance often does not correlate with word frequency but strongly correlates with domain similarities between train and test sets.

In isomorphic settings, Fig. 1 (c) shows that linear alignments are the winner, both in supervised (Rotation) or unsupervised (MUSE) settings. It means a simple linear transformation perfectly aligns vector spaces of isomorphic languages, especially for Rotation. We mark this as a sanity test, as most of languages seem non-isomorphic (Søgaard et al., 2018).

In non-isomorphic settings, Fig. 1 (d) shows that non-linear alignments, both supervised (Real-NVP) and unsupervised (GAN-Real-NVP), are better than linear counterparts when train and test domains are similar; however, linear alignments are indeed better in terms of generalization when train and test domains are dissimilar. One may not forget: many works (Mikolov et al., 2013; Glavaš et al., 2019) conclude that non-linear alignments for static embeddings are always worse than linear ones. We complement this analysis with ours, showing in which cases non-linear alignments based on contextual embeddings succeed and fail. Besides, our findings hint at real scenarios in which linear alignments might be a good choice. For instance, one aligns vector spaces of Modern-English and Simplified-Chinese but evaluates on Middle-English and Traditional-Chinese.

Overall, we show that alignment for static and contextual embeddings have different conclusions. By contrasting them, we offer better understanding of each. For instance, non-linear alignments based on contextual embeddings are good in general, but they have trouble in generalization.

Maximum capacities of unsupervised approaches. Fig. 2 (left) shows how well two languages are aligned when train and test domains are similar. Real-NVP and GAN-Real-NVP perform best according to task performances in corresponding supervised and unsupervised settings, and their resulting vector spaces are better overlapped than others. This confirms the effect of our density based approaches. However, there are big performance gaps between supervised and unsupervised approaches, especially for Real-NVP (73.3) vs. GAN-Real-NVP (41.0), notwithstanding small subspaces overlaps and small differences in model architectures. This confirms that density matching alone is not sufficient. Fig. 2 (right) shows that after bootstrapping GAN-Real-NVP rivals (sometimes exceeds) its supervised variant Real-NVP. We also see similar results from Rotation vs. MUSE. Cross-correlation helps best in all cases, while graph-structure and Procrustes lead to less consistent gains across approaches.

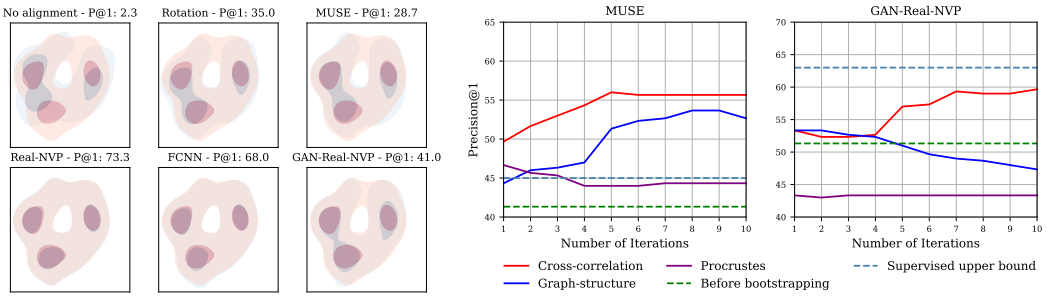

Figure 2: (left) shows how well two languages are aligned according to a visual introspection (subspace overlaps) and Precision@1; (right) compares unsupervised approaches with their supervised variants (Rotation and Real-NVP) in non-isomorphic settings ($t = 5$) as use case. We set the number of occurrences k per word to 100. We evaluate maximum capacities of unsupervised approaches and boostrapping procedures when train and test domain distance ϵ is almost zero—no generalization issues for non-linear approaches. To do so, we directly split data (70/30) to train and test sets.

Overall, these experiments show that bootstrapping procedures are crucial for unsupervised alignments to compete with their supervised variants.

4.4 Experiments on Real Data

XTREME (Hu et al., 2020) and many others are popular benchmarks for evaluating multilingual representations; however, they do not address alignments themselves, but rather focus on how multilingual representations impact cross-lingual systems—as such, the impact of better alignments is indirect and may be lost in complex supervised downstream task classifiers. In this work, we systematically evaluate inherent strengths of alignments through the lens of (i) intrinsic alignment results and (ii) how alignments extrinsically impact multilingual representations without supervised classifiers on top. We leave performances in supervised downstream tasks to future work. Our tasks are listed below:

- CBLI[4] is the contextualized extension of BLI. Both contain a bilingual lexicon per language pair, but CBLI marks each occurrence of a word as an entry in lexicons. For each language pair, we extract 10k word translations from parallel sentences using FastAlign (Dyer et al., 2013). We report Precision@1.
- Alignment (Align) is a bilingual word retrieval task. Each language pair contains gold standard 2.5k word translations annotated from parallel sentences. We use a word alignment tool SimAlign (Jalili Sabet et al., 2020) to retrieve word translations from parallel sentences based on their embeddings. We report F-score that combines precision and recall.
- Reference-free evaluation (RFEval) measures correlations between human and automatic judgments of translation quality. We use XMover (Zhao et al., 2019; 2020b) (excluding a target-side language model) to rate translation quality. It compares system translations with source sentences based on their embeddings. We report Pearson correlations between XMover and human scores. Each language pair covers 3k source sentences, each containing multiple system translations and human judgments.
- Tatoeba is a bilingual sentence retrieval task taken from XTREME. Each language pair contains 1k sentence translations. Given a source sentence, we find its nearest translation from a pool of candidates based on cosine similarities of embeddings. We report Precision@1.

Tatoeba, CBLI and RFEval contain 6 languages German, Czech, Latvian, Finnish, Russian and Turkish, paired to English, while Align contains German/Czech-to-English, as the remaining ones are not available. We investigate two popular multilingual representations: m-BERT and XLM-R. We mark CBLI as an intrinsic task and the others as extrinsic tasks, as CBLI is the only one following the same data creation procedure of our train data (see Setup below).

[4]We provide two complementary CBLI data: one is gold but simulated; another is real but contains noises.

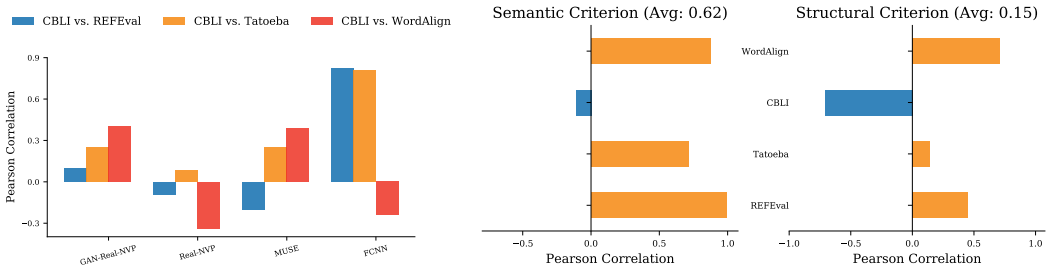

Figure 3: Pearson correlations between intra-task results (left) and between criterion scores and task results from Real-NVP as use case (middle)+(right). For each approach, we run 20 epochs (leading to 20 model candidates), according to which we obtain a list of results in each task and a list of criterion scores used to compute correlations. Results are averaged across language pairs and encoders.

Setup. To contrast supervised and unsupervised approaches, we consider two data scenarios: (i) very limited parallel data and (ii) no parallel data. In case (i), we use 20k (compared to 250k often used in previous work) parallel sentences sampled from News-Commentary (Tiedemann, 2012) for Russian/Turkish-to-English and from EuroParl (Koehn, 2005) for other languages. We extract word translations from parallel sentences using FastAlign, from which we induce permutation matrices as cross-lingual supervision. In case (ii), we unpair word translations obtained from (i) by removing the use of permutation matrices. Thus, we compare supervised and unsupervised approaches under similar conditions, *viz.*, with similar scale of data.

	m-BERT				XLM-R			
Alignments	RFEval	Tatoeba	CBLI	Align	RFEval	Tatoeba	CBLI	Align
Original	27.23	49.35	50.9	61.54	26.42	63.40	48.58	59.77
Supervised mapping functions								
Rotation-20k	38.73	55.28	58.45	**62.83**	34.67	68.60	53.67	60.85
FCNN-20k (semantic criterion)	**42.72**	61.18	55.30	61.50	36.67	**80.20**	50.88	59.87
FCNN-20k (5 epochs)	38.40	58.97	54.02	61.02	33.50	77.98	50.48	59.50
Real-NVP-20k (semantic criterion)	**42.32**	**62.87**	57.62	**62.59**	**44.17**	**80.08**	**61.63**	**62.84**
Real-NVP (5 epochs)	40.12	60.70	58.52	61.80	42.24	78.75	61.76	61.20
GBDD-20k	28.77	52.28	51.42	61.71	27.13	68.85	48.42	59.81
Joint-Align-100k (3 epochs)	41.23	59.13	**64.67**	**62.30**	-	-	-	-
InfoXLM-42GB (150K training steps)	-	-	-	-	37.60	76.10	60.80	**62.94**
Unsupervised mapping functions								
Normalization	30.08	61.28	54.88	**62.54**	39.52	79.75	59.03	62.55
VecMap-20k (∼500 epochs)	30.77	55.00	**64.42**	**62.50**	-	-	-	-
MUSE-20k (5 epochs)	29.20	50.20	52.30	61.56	25.21	63.42	50.20	60.20
MUSE-20k (semantic criterion)	31.23	51.42	52.48	61.64	27.55	65.72	50.00	60.01
+ Cross-Correlation	35.25	52.90	52.87	**62.63**	32.05	69.23	49.80	60.49
+ Graph Structure	33.22	51.65	53.10	62.17	29.48	68.33	50.18	60.46
+ Procrustes	36.85	54.13	55.22	**62.71**	33.37	68.82	50.37	60.59
GAN-Real-NVP-20k (5 epochs)	32.24	59.10	56.79	61.80	39.61	77.77	60.83	60.90
GAN-Real-NVP-20k (semantic criterion)	33.90	61.20	57.03	**62.33**	41.72	79.67	**61.00**	**62.81**
+ Cross-Correlation	35.33	**62.32**	58.00	**62.70**	**42.60**	**80.50**	**61.23**	**63.15**
+ Graph Structure	34.32	61.82	56.65	**62.52**	41.55	80.02	60.83	**62.99**
+ Procrustes	**36.93**	53.95	56.05	**62.79**	33.78	67.95	51.60	60.51

Table 1: Performances are averaged across language pairs. We bold numbers that significantly outperform others according to paired t-test (Fisher, 1935). Joint-Align uses 100k parallel data per language pair; others only use 20k parallel data. InfoXLM uses 42GB parallel data in total but lacks Czech/Latvian/Finish-to-English. We exclude validation criteria for Rotation, GBDD and Normalization, as they have closed-form solutions. VecMap needs ∼500 epochs to reach convergence.

How to select the best model. We compare two choices for model selection during training: using (i) intrinsic data (CBLI) and (ii) our unsupervised validation criteria. Fig. 3 (left) shows that CBLI

results often correlate poorly (or even negatively) with results on other tasks in both supervised and unsupervised settings. This means intrinsic data is a bad choice for model selection. Tab. 2 (appendix) reports correlation statistics across approaches, showing that task performances correlate negatively in about 30% setups. This means when selecting the best model, one should take all these task results into account. We take Real-NVP as use case, to show how to choose our criteria. Fig. 3 (middle)+(right) shows that semantic criterion correlates better than structural criterion with task performances on average (0.62 vs. 0.15). This suggests to use semantic criterion for Real-NVP. We also see similar results for other approaches. Our initial study finds linearly combining two criteria leads to small gains. Thus, we use the semantic criterion for simplicity in all setups.

Overall, these results show that we cannot use intrinsic data for model selection, and when choosing a validation criterion, we should take into account all task performances.

Results. Tab. 1 shows results of approaches on real data in supervised and unsupervised settings. In **supervised** settings, FCNN and Real-NVP training for 5 epochs are worse than their variants that use our criteria for model selection in almost all tasks, especially on RFEval. This confirms the importance of validation criteria. Joint-Align that uses 100k parallel data performs best on intrinsic CBLI; however, in the remaining tasks, it is worse than Real-NVP-20k that uses merely 20k data and our criteria accounting for all task results; we see similar results in unsupervised settings (VecMap vs. GAN-Real-NVP). This means our validation criteria prevent approaches from overfitting intrinsic data. Note that gains on Align are much smaller than on others. Thus, either the word alignment tool SimAlign or the dataset is not sensitive to improved embeddings. This indicates that one should improve evaluation setups such as tools and datasets. Our approach Real-NVP is the strongest approach. It helps for both m-BERT and XLM-R. When using validation criteria, Real-NVP performs best in 7 out of 8 tasks in supervised settings. Tab. 3 (appendix) also confirms the effect of Real-NVP when comparing InfoXLM with our approaches under fair conditions—evaluated across three language pairs for which InfoXLM use parallel data.

In **unsupervised** settings, our criteria are also crucial, as MUSE and GAN-Real-NVP that use our criteria largely outperform training for 5 epochs. Unlike in simulation, there are small performance gaps between supervised and unsupervised approaches, especially the gap between Real-NVP and GAN-Real-NVP (9.0 vs. 32.3 points in simulation). Therefore, while boostrapping procedures offer less help in our real data setups, cross-correlation based boostrapping still helps best for both unsupervised approaches. Procrustes results are similar as in simulation—it only helps for MUSE and harms GAN-Real-NVP. Our GAN-Real-NVP that uses our criteria and cross-correlation based boostrapping performs best in all cases. It rivals the best supervised approach Real-NVP.

Overall, these results show our semantic criterion is helpful; density based approaches are effective in both supervised and unsupervised settings. Unsupervised approaches after bootstrapping based on cross-correlation rival supervised ones. Full results are provided in Tab. 4 (appendix).

5 Conclusion

Resource and typology disparities are primary characteristics of languages, leading to multilingual representations with capabilities unequal across languages. While recent developments on contextualized alignments allow for producing language-agnostic representations, their applications are limited in scope due to data scarcity in low-resource languages. In this work, we demonstrated that our solutions are particularly effective to mitigate the data scarcity issue. With 20k parallel data, our alignments outperformed others that have been trained on 100k parallel data. Further, we showed that parallel data could be removed when integrating unsupervised alignments in boostrapping procedures.

As a form of domain adaptation techniques, density based approaches have been applied in many cross-domain applications, such as image-captioning (Mahajan et al., 2020), image-to-image translation (Grover et al., 2020; Gong et al., 2019), static embeddings (Zhou et al., 2019) and machine translation (Setiawan et al., 2020). In this work, we showed that density based approaches overfit intrinsic data when not trained properly; simple linear techniques indeed perform better when train and test domains are dissimilar. Further, we showed that boostrapping procedures and validation criteria are crucial for density based approaches to improve their effectiveness. While these analyses have been provided in contextualized alignments as use case of cross-domain applications, we believe our findings provide strong cues towards effective domain adaptation techniques in other applications.

REFERENCES

Hanan Aldarmaki and Mona Diab. Context-aware cross-lingual mapping. 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)*, 2019.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018a. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018b. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7375–7388, Online, July 2020. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*, 2020.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3576–3588, Online, June 2021. 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. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. 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)*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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)*, Hong Kong, China, November 2019. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*, 2014.

RA Fisher. The design of experiments (oliver and boyd, edinburgh, london). 1935.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 710–721, Florence, Italy, July 2019. Association for Computational Linguistics.

Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Aditya Grover, Christopher Chute, Rui Shu, Zhangjie Cao, and Stefano Ermon. Alignflow: Cycle consistent learning from multiple domains via normalizing flows. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 13–18 Jul 2020.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3633–3643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.284. URL https://aclanthology.org/2021.naacl-main.284.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5. Citeseer, 2005.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, November 2020.

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

Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. In *International Conference on Learning Representations*, 2020.

Xia Mengzhou, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Newbig, and Ahmed Hassan Awadallah. Metaxl: Meta representation transformation for low-resource cross-lingual learning. 2021.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013.

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

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, Lille, France, 07–09 Jul 2015. PMLR.

Hendra Setiawan, Matthias Sperber, Udhyakumar Nallasamy, and Matthias Paulik. Variational neural machine translation with normalizing flows. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

Shijie Wu and Mark Dredze. 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)*, Hong Kong, China, November 2019. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2020. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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)*, Hong Kong, China, November 2019. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. Inducing language-agnostic multilingual representations. *CoRR*, abs/2008.09112, 2020a.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020b.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. Density matching for bilingual word embedding. 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)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Settings	Alignments	$[-1,0]$	$(0,0.4]$	$(0.4,1]$
Supervised	FCNN-20k	50%	0%	50%
	Real-NVP-20k	33%	17%	50%
Unsupervised	MUSE-20k	17%	66%	17%
	GAN-Real-NVP-20k	17%	50%	33%

Table 2: Correlation statistics: last three columns denote intervals used to split Pearson's ρ range. Each entry denotes the percent of all task pairs that have ρ in corresponding intervals. For instance, 17%-50% of task pairs across approaches have negative correlations. We run 20 epochs to obtain a list of results in each task used to compute correlations. Results are averaged across language pairs and encoders.

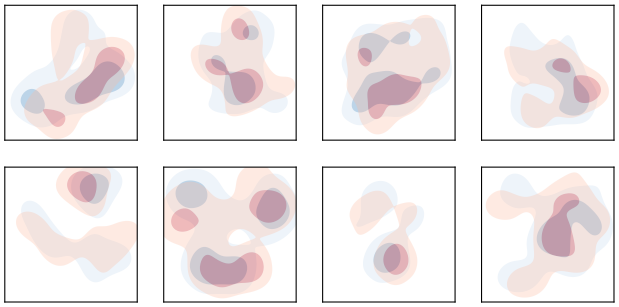

Figure 4: Eight figures are produced by simulation. Each contains a language pair with two outlined subspaces, colored in blue and red. Each subspace contains 1-3 areas merely for simplicity, each representing a word—whereas our simulation results are based on 20 words. Each area (or each word) contains a few contextual embeddings sampled from a two-dimensional Gaussian distribution.

A Appendix

A.1 Connection with canonical correlation and language isomorphism.

Cross-correlation between random vectors are often measured via Canonical Correlation Analysis (CCA). While Faruqui & Dyer (2014) find CCA is useful to improve static embeddings, it requires finding k primary canonical variables. However, our solution is simple. It measures cross-correlation using random vectors per se.

In graph theory, two graphs are called isomorphic if corresponding adjacency matrices $\mathbf{A}$ and $\mathbf{B}$ are permutation similar, denoted by $\mathbf{PAP}^{-1} = \mathbf{B}$ with $\mathbf{P}$ as a permutation matrix. According to Eq. 4, $\mathbf{A}$ and $\mathbf{B}$ are denoted by $\mathbf{M}_{\tilde{Y}}\mathbf{M}_{\tilde{Y}}^{\intercal}$ and $\mathbf{M}_{Y}\mathbf{M}_{Y}^{\intercal}$; minimizing the difference between $\mathbf{A}$ and $\mathbf{B}$ ideally leads to $\mathbf{P}$ as an identity matrix when converged. As such, applying graph-structure approach meets the condition of graph isomorphism—which has been profiled as a form of language isomorphism (Søgaard et al., 2018). Therefore, our approach theoretically yields isomorphic multilingual subspaces for non-isomorphic (distant) languages.

	RFEval			Tatoeba			CBLI			Align
Alignments	DE-EN	RU-EN	TR-EN	DE-EN	RU-EN	TR-EN	DE-EN	RU-EN	TR-EN	DE-EN
Original	25.1	21.4	36.5	89.0	70.2	55.4	54.1	60.9	47.9	78.5
Real-NVP-20k	**40.5**	**38.0**	**54.7**	**94.9**	83.9	75.6	64.6	**71.7**	**63.9**	80.5
InfoXLM-42GB	37.3	34.9	52.6	**95.5**	**85.7**	**86.1**	**67.3**	**72.4**	**64.0**	**82.6**

Table 3: Comparison of XLM-R (original) and its variant after alignments. InfoXLM-42GB has been trained for 150K training steps, while Real-NVP-20k uses semantic criterion for model selection. We bold numbers that significantly outperform others according to paired t-test (Fisher, 1935). Real-NVP that uses our criteria performs comparable with InfoXLM; the latter uses much larger parallel data but lacks validation criteria. This confirms the effectiveness of our approaches.

A.2 TRAINING DETAILS

For Real-NVP and GAN-Real-NVP, the Real-NVP component is stacked with two FCNN networks with a dimension size of 512 for each. We optimize them with RMSprop and 32 batch size, learning rate 1e-4 for Real-NVP and learning rate 1e-6 for GAN-Real-NVP. The coefficient for maximum likelihood estimation is set to 1e-4. For FCNN, we set the network with 1024 as a dimension size. We optimize it with Adam, learning rate 1e-4 and 32 batch size. For MUSE, we optimize it with SGD, learning rate 1e-1 and 32 batch size. For Joint-Align, we follow the parameters in Cao et al. (2020). We run all the baselines and our models on a single GTX 1070 for synthetic data, but run them on a single Tesla V100 for real data. We align the 8-th layer of multilingual representations such as m-BERT and XLM-R for individual language pairs, as previous works (Ethayarajh, 2019; Zhao et al., 2020a) show that the representations obtained from upper layers often have poor discriminativeness, i.e., a failure to distinguish mutual word translations from random word pairs.

Lang	Original	Rotation	FCNN	NVP	GBDD	Joint-Align	NORM	MUSE	MUSE+C	MUSE+G	MUSE+P	GNVP	GNVP+C	GNVP+G	GNVP+P	infoXLM	VecMap
Reference-free Evaluation (m-BERT)																	
cs-en	23.6	32.5	41.5	30.2	24.4	32.5	22.5	24.8	29.7	26.7	30.4	24.9	25.3	25.2	30.5	-	25.0
de-en	29.8	31.6	34.1	34.2	30.0	39.9	31.4	29.9	31.3	31.1	31.2	32.8	33.0	32.8	31.3	-	31.8
fi-en	30.9	48.5	58.3	51.8	32.5	48.6	32.3	37.9	44.0	40.4	46.0	37.2	39.5	37.9	46.1	-	39.1
lv-en	23.0	44.0	55.1	52.1	25.7	46.1	31.3	33.4	35.9	33.5	40.9	38.0	40.7	38.8	41.0	-	24.9
ru-en	19.4	23.4	30.5	33.4	21.4	31.7	23.9	19.8	22.0	21.3	22.9	26.5	27.0	26.3	22.9	-	18.7
tr-en	36.7	52.4	60.8	52.2	38.6	48.6	39.1	41.6	48.6	46.3	49.7	44.0	46.5	44.9	49.8	-	44.5
Reference-free Evaluation (XLM-R)																	
cs-en	20.6	26.3	26.1	32.9	22.4	-	28.0	20.4	24.2	22.3	24.6	29.2	30.1	29.1	25.1	26.6	
de-en	25.1	27.4	28.2	40.5	25.4	-	39.5	25.2	26.5	26.0	26.6	40.4	40.4	40.5	27.2	37.3	
fi-en	26.6	39.9	44.6	47.7	26.0	-	41.6	29.0	35.7	32.3	37.8	44.3	47.2	44.5	38.2	36.1	
lv-en	28.3	42.3	48.6	51.2	29.6	-	44.1	30.8	37.9	32.7	40.7	47.3	48.7	47.4	40.9	38.3	
ru-en	21.4	25.6	27.7	38.0	21.5	-	34.3	21.5	23.3	22.4	24.6	35.9	36.3	36.1	25.1	34.9	
tr-en	36.5	46.5	44.8	54.7	37.9	-	49.6	38.4	44.7	41.2	45.9	53.2	52.9	51.7	46.2	52.6	
Tatoeba (m-BERT)																	
cs-en	47.5	53.1	61.4	61.3	48.3	51.1	60.6	48.2	50.8	48.7	51.6	60.6	61.3	60.7	51.5	-	47.5
de-en	76.6	79.5	84.5	87.3	78.6	89.6	86.0	77.7	79.3	78.7	78.9	86.2	86.5	86.0	78.8	-	81.2
fi-en	41.5	50.6	57.2	54.3	46.4	70.4	53.5	44.7	46.5	45.2	47.8	53.9	56.1	54.6	47.5	-	53.5
lv-en	31.7	39.8	47.4	51.2	35.0	31.1	44.0	35.3	35.0	33.8	37.1	44.1	44.7	43.8	37.0	-	33.5
ru-en	63.0	65.7	70.3	72.6	64.8	74.2	75.4	63.5	64.9	64.7	66.3	75.1	76.0	75.1	66.3	-	66.5
tr-en	35.8	43.0	46.3	50.5	40.6	38.4	48.2	39.1	40.9	38.8	43.1	47.3	49.3	48.1	42.6	-	47.6
Tatoeba (XLM-R)																	
cs-en	49.8	57.5	74.6	72.9	59.2	-	72.5	53.8	58.9	60.6	59.1	72.8	74.0	73.3	57.6	69.6	
de-en	89.0	91.0	94.5	94.9	90.1	-	95.2	88.7	91.0	89.6	90.8	94.6	95.2	94.8	90.5	95.5	
fi-en	63.8	68.6	81.3	80.8	69.3	-	79.6	65.6	69.8	67.2	68.8	80.6	81.2	80.6	67.7	68.4	
lv-en	52.2	61.3	72.6	72.4	58.9	-	71.9	56.4	63.1	59.2	61.0	71.5	71.5	71.7	59.7	51.2	
ru-en	70.2	71.5	83.1	83.9	73.2	-	83.9	71.2	70.6	70.2	72.2	83.4	84.2	84.0	70.9	85.7	
tr-en	55.4	61.7	75.1	75.6	62.4	-	75.4	58.6	62.0	63.2	61.0	75.1	76.9	75.7	61.3	86.1	
Contextual BLI (m-BERT)																	
cs-en	47.3	52.6	49.2	52.3	47.7	57.0	50.1	47.7	50.0	48.4	50.8	51.8	52.0	51.5	51.0	-	55.1
de-en	61.0	64.6	59.0	63.8	60.9	79.4	63.3	61.8	61.8	62.1	62.5	64.8	64.8	64.7	63.0	-	71.7
fi-en	36.8	48.9	45.8	45.8	37.9	74.2	43.0	40.7	40.6	40.7	43.4	45.0	47.6	44.7	44.6	-	56.7
lv-en	42.7	51.8	49.5	50.6	43.5	49.8	47.9	45.5	44.5	45.9	48.7	49.9	50.5	49.7	49.5	-	53.9
ru-en	66.5	70.9	67.6	71.4	67.0	67.4	69.2	67.2	67.5	67.8	69.5	70.5	71.0	70.6	69.9	-	76.9
tr-en	51.1	61.9	60.7	61.8	51.5	60.2	55.8	52.0	52.8	53.7	56.4	60.2	62.1	58.7	58.3	-	70.4
Contextual BLI (XLM-R)																	
cs-en	40.2	45.1	43.6	51.1	39.3	-	48.2	41.6	42.0	41.5	42.3	49.8	50.0	49.6	43.3	50.7	
de-en	54.1	57.9	53.2	64.6	54.3	-	62.7	55.2	55.1	55.3	55.4	65.1	65.3	64.9	56.3	67.3	
fi-en	44.2	51.0	46.1	60.9	44.0	-	58.1	46.8	45.7	47.2	46.6	60.2	60.5	59.9	48.3	57.1	
lv-en	44.2	50.0	48.3	57.6	44.2	-	54.8	46.4	45.7	47.3	46.5	56.9	57	56.7	47.8	53.5	
ru-en	60.9	64.4	60.3	71.7	60.7	-	70.3	61.4	61.8	61.2	62.2	71.5	71.5	71.4	63.0	72.4	
tr-en	47.9	53.6	53.8	63.9	48.0	-	60.1	48.6	48.5	48.6	49.2	62.5	63.1	62.5	50.9	64.0	
Word Alignment (m-BERT)																	
cs-en	44.0	45.0	43.4	45.7	44.1	44.2	44.6	44.0	45.0	44.5	45.0	44.6	45.1	44.9	45.1	-	45.0
de-en	79.1	80.6	79.6	79.5	79.3	80.2	80.5	79.3	80.2	79.9	80.4	80.0	80.3	80.2	80.5	-	80.0
Word Alignment (XLM-R)																	
cs-en	41.1	42.1	41.8	45.2	40.9	-	44.4	41.2	42.1	41.8	42.1	44.6	44.9	44.8	42.0	43.3	
de-en	78.5	79.6	77.9	80.5	78.7	-	80.8	78.8	78.9	79.2	79.1	81.0	81.4	81.2	79.0	82.6	

Table 4: Full results of baselines and our alignments. Real-NVP and GAN-Real-NVP are denoted by NVP and GNVP. [Method]+P/C/G denotes the integration in the Procrustes refinement (P), or in our bootstrapping procedure constrained with cross-correlation (C) and with graph structure (G).