UNSUPERVISED HYPER-ALIGNMENT FOR MULTILINGUAL WORD EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the problem of aligning continuous word representations, learned in multiple languages, to a common space. It was recently shown that, in the case of two languages, it is possible to learn such a mapping without supervision. In this paper, we extend one of the proposed methods to the problem of aligning multiple languages to a common space. A simple solution to this problem is to independently map all languages to English. Unfortunately, this can degrade the alignments between languages different than English. We thus propose to add constraints to ensure that the learned mappings can be composed, leading to better alignments. We evaluate our method on the problem of aligning word vectors in eleven languages, showing improvement in word translation requiring the composition of multiple mappings.

1 INTRODUCTION

Pre-trained continuous representations of words are standard building blocks of many natural language processing and machine learning systems (Mikolov et al., 2013b). Word vectors are designed to summarize and quantify semantic nuances through a few hundred coordinates. Such representations are typically used in challenging down-stream tasks to improve generalization when the amount of data is scarce (Collobert et al., 2011). The distributional information used to learn these word vectors derives from statistical properties of word co-occurrence found in large corpora (Deerwester et al., 1990). Such corpora are, by design, monolingual (Mikolov et al., 2013b; Bojanowski et al., 2016), resulting in the independent estimation of word embeddings for each language.

A limitation of these monolingual embeddings is that it is impossible to compare words across languages. It is thus tempting to combine all these word representations into a common multilingual space, where every language could be mapped. Mikolov et al. (2013a) noticed that word vectors learned on different languages share a similar structure. More precisely, two sets of pre-trained vectors in different languages can be aligned to some extent: a linear mapping between the two sets of embeddings is enough to produce decent word translations. Recently, there has been an increasing interest in mapping these pre-trained vectors in a common space (Xing et al., 2015b; Artetxe et al., 2017), resulting in many publicly available embeddings in many languages mapped into a single common vector space (Smith et al., 2017; Conneau et al., 2017; Joulin et al., 2018). The quality of these multilingual embeddings can be tested by composing mappings between languages and looking at the resulting translations. Figure 1 shows the quality of word translation between English and Italian as a function of the number of composed mappings used between these two languages. The linear degradation of their quality is an empirical evidence that independent alignments of language pairs do not guarantee a "universal" coherence of the embedding space among languages. Practically speaking, it is not surprising to see such a degradation since these bilingual alignments are trained separately, without enforcing transitivity.

We propose a novel approach to align multiple languages simultaneously in a common space in a way that enforces transitive translations. Our method relies on constraining mappings to verify circular translations over a graph of languages. Nakashole and Flauger (2017) has recently shown that such constraints over a well chosen triplet of languages improve supervised bilingual alignment. We extend their work to an unsupervised multilingual setting, i.e., where no bilingual lexicon nor language graph is provided. We show that our approach achieves competitive performance among unsupervised approaches while enforcing composition.

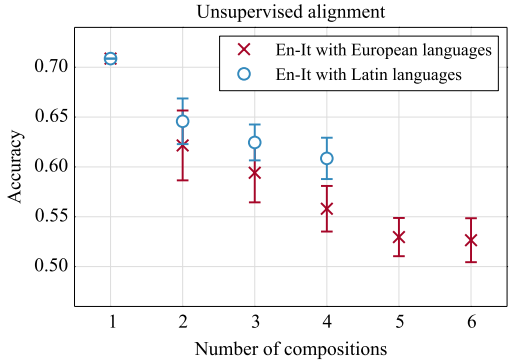


Figure 1: English to Italian word translation when mapping through an increasing number k of intermediate languages. For instance, a possible path is English$\rightarrow$ French$\rightarrow$ Russian$\rightarrow$ Italian, where the English vectors are transformed by applying successively three mappings $Q_{\text{en-fr}}$, $Q_{\text{fr-ru}}$, and $Q_{\text{ru-it}}$. The accuracy is averaged over many paths. We align the fastText embeddings of Bojanowski et al. (2016) with the approach of Grave et al. (2018) and a Nearest Neighbor (NN) criterion.

2 Preliminaries on bilingual alignment

In this section, we provide a brief overview on bilingual alignment approaches to map a dictionary of embeddings onto another, and discuss their shortcomings when used in a multilingual setting.

2.1 Supervised bilingual alignment

Mikolov et al. (2013a) formulate the problem of word embedding alignment as an optimization problem with a quadratic loss. Given two aligned sets of (row) word vectors stacked in two $n \times d$ matrices $X = [x_1; \ldots; x_n]$ and $Y = [y_1; \ldots; y_n]$ respectively, we learn a linear mapping matrix Q by solving the following least-square problem

$$\min_{Q \in \mathbb{R}^{d \times d}} \|XQ - Y\|_2^2,$$

which admits a closed form solution. A crucial assumption of that formulation is that each embedding (each line) in X corresponds to a word whose translation has an embedding stored in Y at the same index. Therefore, a bilingual lexicon is needed to form Y and X. Xing et al. (2015b) have shown that constraining the mapping Q to the set of orthogonal matrices $\mathbb{O}_d$ leads to better transformations, while preserving the dot product between vectors from the same language. As shown by Schnemann (1966), this reformulation, commonly known as Orthogonal Procrustes, has a closed form solution equal to $Q = UV^\top$, where $USV^\top$ is the singular value decomposition of $X^\top Y$.

2.2 Unsupervised bilingual Alignment: Wasserstein-Procrustes

Wasserstein-Procrustes (Zhang et al., 2017a; Grave et al., 2018) extends orthogonal Procrustes to the setting of unsupervised bilingual alignment. In this approach, an orthogonal transformation and a one-to-one word mapping are jointly learned. This mapping is parametrized using a $n \times n$ binary matrix P, belonging to the set of assignment matrices $\mathcal{P}_n$, itself consisting in all bistochastic matrices (the Birkhoff polytope) with binary entries:

$$\mathcal{P}_n = \mathcal{B}_n \cap \{0,1\}^{n \times n}, \text{ where } \mathcal{B}_n = \left\{ P \in \mathbf{R}_+^{n \times n}, P\mathbf{1}_n = \mathbf{1}_n,\ P^\top \mathbf{1}_n = \mathbf{1}_n \right\}.$$

The joint optimization leads to the following problem:

$$\min_{Q \in \mathcal{O}_d} \min_{P \in \mathcal{P}_n} \|XQ - PY\|_2^2. \tag{1}$$

This problem is not convex since neither the feasible set $\mathcal{P}_n$ nor $\mathcal{O}_d$ is convex. Treating each variable separately leads, however, to two well understood optimization problems: when P is fixed, solving for Q, involves solving the orthogonal Procrustes problem. When Q is fixed, an optimal permutation matrix P can be obtained with the Hungarian algorithm. Both algorithms have a cubic complexity, although on different quantities: Procrustes involves the dimension of the vectors (typically $d = 300$) whereas the Hungarian algorithm is applied on the vocabulary (typically $N = 20$k–200k). Because these two problems can be easily solved, a simple heuristic is to address Eq.(1) using alternate optimization.

We reformulate the optimal matching as a network flow problem using the Kantorovich relaxation of optimal transport (Villani, 2003; Santambrogio, 2015) on the Birkhoff polytope

$$\min_{P \in \mathcal{B}_n} \langle P, D^2_{XQ,Y} \rangle, \tag{2}$$

where $D^2_{XQ,Y}$ is the $n \times n$ matrix with entries $\|x_i Q - y_j\|_2^2$. This quantity is well known as the *2-Wasserstein* distance between the two point clouds described in XQ and Y. Cuturi (2013) proposed to regularize problem Eq.(2) with the negative entropy of P. The resulting problem is strongly convex and a solution can be obtained with a complexity of $O\left(N^2\right)$ (from matrix vector multiplications). Grave et al. (2018) reduces further the overall complexity by using a stochastic optimization.

As is usually the case for many non-convex problems, a good initial guess can help converge to better local minima. Grave et al. (2018) propose for instance to compute an initial P by solving a convex relaxation of the quadratic assignment problem which can ideally disambiguate among several possible symmetric choices. We found, however, that a different approach, grounded on the entropic regularization of the Gromov-Wasserstein (GW) problem (Mémoli, 2011) worked well in practice and was significantly faster (see also Alvarez-Melis and Jaakkola (2018) for a recent use of GW in a NLP context). As described in Solomon et al. (2016); Peyré et al. (2016), the entropic regularization of GW consists in computing an optimal assignment matrix $\hat{P}$ by solving

$$\min_{P \in \mathcal{P}_n} \langle -D^2_X P D^2_Y, P \rangle - \varepsilon \mathcal{H}(P), \tag{3}$$

where $\mathcal{H}(P)$ stands for the Shannon entropy of P and D^2_X, D^2_Y for the $n \times n$ pairwise squared Euclidian distance matrices of all points in X and Y respectively. Note that the case $\varepsilon = 0$ corresponds to Mémoli's initial proposal. Optimizing the regularized version ($\epsilon > 0$) leads to a local minimum $\hat{P}$ that can be used as an initialization to solve Eq. (1).

3 Composable Multilingual Alignements

Most publicly available aligned vectors are mapped to a unique multilingual vector space, that typically coincides with English (Smith et al., 2017; Conneau et al., 2017; Joulin et al., 2018). This means that any translation to English can be carried out with a single matrix multiply, but that any translation between two other languages requires two matrix multiplications (to, and then from English). The results presented in Figure 1 are evidence that these aligned vectors would lead to degraded performance when translating or comparing words from two languages different than English, compared to directly learning a mapping between these two languages. In this section, we propose a method to jointly align N sets of word embeddings to a unique common space while preserving the quality of word translations between all pairs of languages. We discuss efficient solutions that avoid learning N^2 matrices.

3.1 Compositionality as a set of triplet constraints

In the rest of this paper, we focus on the unsupervised multilingual alignment problem, that can be formulated in the following way: given a graph $\mathcal{G} = (V, E)$ where each node is a language i associated with its embeddings X_i, we are interested in learning the translation matrix Q_{ij} for each edge (i, j) in E, i.e., learning a mapping between each pair of languages connected in the graph. This graph can be dense, requiring to learn N^2 matrices, or can be a star tree with a given language in the middle (e.g., English) and only N matrices. In the unsupervised setting, we also need to learn a permutation matrix P_{ij} for each edge, leading to the following optimization problem:

$$\min_{Q_{ij} \in \mathcal{O}_d,\ P_{ij} \in \mathcal{P}_n} \quad \sum_{(i,j) \in E} \|X_i Q_{ij} - P_{ij} X_j\|_2^2. \tag{4}$$

This loss naturally decomposes along edges, leading to $|E|$ independent bilingual alignment problems and no guarantees of compositionality. Nakashole and Flauger (2017) suggest to constrain bilingual alignments along a triplet of languages, improving, in the supervised setting, the alignment of distant languages by taking a language "in-between". In absence of supervision or knowledge about similarities between languages, we can still adapt this idea by adding circular constraints

Algorithm 1 Language Tree derivation

Input: embedding sets $(X_1, \cdots, X_N)$ **of size** b**; mappings** $(Q_1, \cdots, Q_N)$
for i=1 to N **do**
  for j=i+1 to N **do**
    $Q_{ij} = Q_i Q_j^\top$
    Compute $d_{i,j} = \min_{P \in \mathcal{P}_b} \|X_i W - P X_j\|_2^2$
Build fully connected graph $\mathcal{G} = (V, E)$ with edge weights $(d_{i,j})_{i,j}$.
Return T a minimum spanning-tree of $\mathcal{G}$

to Eq. (4), i.e., constraining mappings over a set $\mathcal{T}$ of triplets of languages to be coherent:

$$\min_{Q_{ij} \in \mathcal{O}_d,\ P_{ij} \in \mathcal{P}_n} \quad \sum_{(i,j) \in E} \|X_i Q_{ij} - P_{ij} X_j\|_2^2 + \mu \sum_{(i,j,k) \in \mathcal{T}} \|Q_{ij} Q_{jk} - Q_{ik}\|_2^2. \tag{5}$$

The set $\mathcal{T}$ must be coherent with the graph $\mathcal{G}$: A constraint (i, j, k) is only considered if the edges (i, j) and (j, k) exist in the graph $\mathcal{G}$. One limitation of this approach is that it potentially requires additional linear mappings corresponding to the constraints. For example, in the case of a tree, these additional edges do not exist as they would create cycles. Instead, we replace the constraints on the mappings by constraints on the mapped vectors, at the cost of additional matchings P_{jk}:

$$\min_{Q_{ij} \in \mathcal{O}_d,\ P_{ij} \in \mathcal{P}_n} \quad \sum_{(i,j) \in E} \|X_i Q_{ij} - P_{ij} X_j\|_2^2 + \mu \sum_{(i,j,k) \in \mathcal{T}} \|X_i Q_{ij} Q_{jk} - P_{ik} X_k\|_2^2. \tag{6}$$

The linear mappings are orthogonal, leading to the following equivalent reformulation:

$$\min_{Q_{ij} \in \mathcal{O}_d,\ P_{ij} \in \mathcal{P}_n} \quad \sum_{(i,j) \in E} \|X_i Q_{ij} - P_{ij} X_j\|_2^2 + \mu \sum_{(i,j,k) \in \mathcal{T}} \|X_i Q_{ij} - P_{ik} X_k Q_{jk}^\top\|_2^2. \tag{7}$$

The general problem stated in Eq. (7) can be optimized with the same alternative minimization procedure as in Wasserstein-Procrustes.

We now discuss how to chose the set of edges E and constraints $\mathcal{T}$. We want to learn only $N - 1$ matrices, while still being able to translate between any two pairs of languages. The set of edges E must thus correspond to a spanning tree. Given a tree E, we can obtain a set of constraints by considering all the triplets (i, j, k) such that the edges (i, j) and (j, k) exists in E. This corresponds to adding constraints between pairs of languages which are at a distance of 2 in the tree.

Minimum spanning tree. Given a set of N^2 residual alignment errors between all pairs of languages, it is possible to compute a minimum spanning tree, where similar languages would be close. In the unsupervised case, we can obtain good enough alignment errors by using the results of our initialization (i.e. bilingual alignment using Wasserstein Procrustes). Figure 2 gives an example of a tree obtained this way. This solution ensures that a translation between two languages follows a path between similar languages, which was shown to work well in practice (Nakashole and Flauger, 2017). However this method is sensitive to the quality of the initialization and translation error can add up over a long path.

Star tree. An alternative to computing a minimum spanning tree is to use a star tree, centered around one language (e.g. English). In that case, every embedding is only one transformation away from a "central" multilingual space. The main advantage of this approach is to limit the length of the path between any pair of language to 2. Unfortunately, this graph leads to $(N - 1)^2$ constraints, since any pair of language (but the central one) is at a distance of 2.

Star tree with pruned constraints. A potential solution to this problem is to limit the number of constraints that are added to our optimization problem. Here we propose to keep only a subset of constraints of size of the order of N. This set can either be random or generated from the minimum spanning tree. It can also be dynamically changed during the optimization. In that way, we have the best of both worlds: our constraints focus only on a subset of languages, and every pair of languages is only at a maximum distance of 2 in the graph. By indexing the central language by 0, this approach

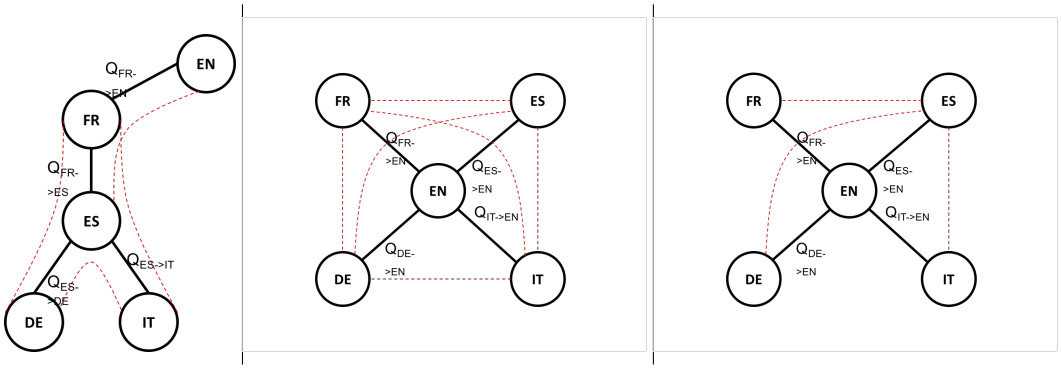


Figure 2: Three alignment models. Plain black lines indicate pairs of languages for which mappings are learned while dashed red lines indicate constraints. **Left:** tree model: alignments are both learned and penalized using a predefined language tree. **Center:** star model: languages are aligned onto a common pivot language. All the N^2 pairs are penalized. **Right:** HUG: languages are aligned onto a common pivot languages but only edges from a predefined language tree are constrained.

Algorithm 2 Hyperalignement with Unsupervised Graphs (HUG)

Input: Embedding sets $(X_0, \cdots, X_N)$; regularization factor μ
for i=1 to N **do** ▷ Initialization
    Q_i=Wasserstein-Procrustes(X_i, X_0)
for epoch=1 to nepoch **do**
    Draw a set of N constraints either randomly or from a tree as constructed with Algo. 1
    Update Q_i by optimizing Eq. (7) with Projected Stochastic Gradient Descent
Return $(Q_1, \cdots, Q_N), T$.

learns a mapping for each language different than 0, leading to the following optimization problem:

$$\min_{Q_i \in \mathcal{O}_d,\ P_{ij} \in \mathcal{P}_n} \sum_{i=1}^{N} \|X_i Q_i - P_{i0} X_0\|_2^2 + \mu \sum_{(i,k)\in\mathcal{T}} \|X_i Q_i - P_{ik} X_k Q_k\|_2^2, \tag{8}$$

where $\mathcal{T}$ is some set of N constraints. If the regularization parameter μ is set to one and we add a mapping Q_0 equal to the identity, we can even rewrite this objective function as:

$$\min_{Q_i \in \mathcal{O}_d,\ P_{ij} \in \mathcal{P}_n} \sum_{(i,j)\in\mathcal{T}'} \|X_i Q_i - P_{ij} X_j Q_j\|_2^2. \tag{9}$$

This problem corresponds to mapping all vectors in a common space, while minimizing the Wasserstein distance between pairs of languages in $\mathcal{T}'$. It does not have any additional hyper-parameters compared to the bilingual method, and we use this formulation in the rest of the paper.

Dynamic constraint tree. So far, we have assumed that the set of triplets is given a priori, but there is no reason to keep it fixed during training. For example, we can sample a new tree at each iteration or learn it from the language distances. Indeed, once all the languages are aligned in a common space, we can compute an optimal 1-to-1 assignment between a pair of languages (i, j) to build a pseudo-distance, $d_{i,j} = \min_{P \in \mathcal{P}_b} \|X_i Q_i - P X_j Q_j\|_2$. Once the distances for all pairs are computed, we compute a minimum spanning tree. Algorithm 1 summarizes this procedure.

Overall approach. Algorithm 2 summarizes our approach, called Hyper-alignment with Unsupervised Graphs (HUG): we initialize the Q_i for each $i = 1, \ldots, N$ with Wasserstein-Procrustes between X_i and the referential X_0. We then iteratively draw N constraints, either randomly or from a tree built on language distances obtained from the alignments. Then we update the alignment matrices by optimizing the problem described in Eq. (7). We refer to our method as HUG-r if the constraints are randomly sampled, and HUG-s if based on a minimum spanning tree.

4 RELATED WORK

Bilingual word embedding alignment. Since the work of Mikolov et al. (2013a), many have proposed different approaches to align word vectors with different degrees of supervision, from fully supervised (Dinu et al., 2014; Xing et al., 2015a; Artetxe et al., 2016; Joulin et al., 2018) to little supervision (Smith et al., 2017; Artetxe et al., 2017) and even fully unsupervised (Zhang et al., 2017a; Conneau et al., 2017; Hoshen and Wolf, 2018). Among unsupervised approaches, some have explicitly formulated this problem as a distribution matching: Cao et al. (2016) align the first two moments of the word vector distributions, assuming Gaussian distributions. Others (Zhang et al., 2017b; Conneau et al., 2017) have used a Generative Adversarial Network framework (Goodfellow et al., 2014). Zhang et al. (2017a) shows that an earth mover distance can be used to refine the alignment obtained from a generative adversarial network, drawing a connection between word embedding alignment and Optimal Transport (OT). Closer to our work, Grave et al. (2018) and Alvarez-Melis and Jaakkola (2018) have proposed an unsupervised bilingual alignment method solely based on OT. We use an approach inspired by their work to initialize our multilingual approach.

Nakashole and Flauger (2017) show that constraining coherent word alignments between triplets of nearby languages improves the quality of induced bilingual lexicons. As opposed to our work, their approach is supervised both in terms of lexicon and choice of triplets. We extend their work in two ways: we jointly extend the graph to more than three languages and we discover it along with the bilingual lexicons without any supervision.

Optimal Transport. Optimal transport (Villani, 2003; Santambrogio, 2015) provides a natural topology on shapes and discrete probability measures (Peyré et al., 2017), that can be leveraged thanks to fast OT problem solvers (Cuturi, 2013; Altschuler et al., 2017). Of particular interest is the Gromov-Wasserstein distance (Gromov, 2007; Mémoli, 2011). It has been used for shape matching under its primitive form (Bronstein et al., 2006; Mémoli, 2007) and under its entropy-regularized form (Solomon et al., 2016).

Hyperalignment. Hyperalignment, as introduced by Goodall (1991), is the method of aligning several shapes onto each other with supervision. Recently, Lorbert and Ramadge (2012) extended this supervised approach to non-Euclidean distances. We recommend Gower et al. (2004) for a thorough survey of the different extensions of Procrustes and to Edelman et al. (1998) for algorithms involving orthogonal constraints. For unsupervised alignment of multiple shapes, Huang et al. (2007) use a pointwise entropy based method and apply it to face alignment.

5 EXPERIMENTAL RESULTS

Implementation Details. We use normalized fastText word embeddings trained on the Wikipedia Corpus (Bojanowski et al., 2016). Word translations are retrieved with a 1-Nearest Neighbor (NN) criterion. We use stochastic gradient descent (SGD) with an initial batch size of 500, 10k iterations per epoch and a fixed learning rate of 0.1. We fix the total number of epoch to 5. At each epoch, we double the batch size and divide by 2 the number of iterations per epoch. We initialize with the Gromov-Wasserstein approach applied to the first 2k vectors and a regularization parameter ε of 0.5 (Peyré et al., 2016). We use the python optimal transport package [1].

Extended MUSE Benchmark. We evaluate our model on the MUSE test datasets (Conneau et al., 2017) on the following languages: Czech, Danish, Dutch, English, French, German, Italian, Polish, Portuguese, Russian and Spanish. MUSE bilingual lexicon are mostly translations to or from English. For missing pairs of languages (e.g., Danish-German), we use the intersection of their translation to English to build a test set. The resulting bilingual lexicons are noisy, but they seem to capture decently the difference between approaches.

Baselines. We consider as baselines several bilingual alignment methods that are either supervised, i.e., Orthogonal Procrustes and RCSLS (Joulin et al., 2018), or unsupervised, i.e., Adversarial (Conneau et al., 2017), ICP (Hoshen and Wolf, 2018) and Wasserstein Procrustes (Grave et al., 2018). We

[1] POT, `https://pot.readthedocs.io/en/stable/`

	Direct			Ind.	Direct			Ind.	Direct			Ind.
	ru-en	pl-en	ru-pl	ru-pl	it-en	pt-en	it-pt	it-pt	nl-en	de-en	nl-de	nl-de
Bil.	42.8	**53.0**	51.7	42.7	**71.4**	**72.5**	74.9	70.3	**64.3**	**66.4**	66.4	63.6
HUG	**44.3**	49.7	-	**50.9**	71.0	71.8	-	**74.0**	63.6	**66.3**	-	**64.8**
	ru-en	pt-en	ru-pt	ru-pt	da-en	cs-en	da-cs	da-cs	fr-en	da-en	fr-da	fr-da
Bil.	42.8	**72.5**	43.7	38.7	49.0	46.3	40.0	33.5	**76.6**	49.0	45.1	41.9
HUG	**44.9**	71.0	-	**43.0**	**50.1**	**47.6**	-	**38.1**	74.0	**49.3**	-	**45.2**

Table 1: Accuracy with a NN criterion on triplet alignment with direct translation ("Direct") and indirect translation ("Ind."). Indirect translation uses English as a pivot language. We consider nearby languages on the first row of results, and distant ones on the second row. For reference, we also provide the performance of direct bilingual alignment. All the accuracies are the average of both directions (source↦target and target↦source). The bilingual baseline ("Bil.") is the Procrustes-Wasserstein of Grave et al. (2018) with a Gromov-Wasserstein initialization.

also consider several alternatives to our approach as baselines: the minimum spanning tree approach ("MST") described on the left panel of Fig. (2). The minimum spanning tree is built on the distances of the Gromov-Wasserstein initialization. We also consider a star-tree with N^2 constraints ("Star").

5.1 Triplet alignment

In this section, we evaluate the quality of our formulation in the simple case of triplets of languages. English is used as a "pivot" for the translation between two other languages. We evaluate both the direct translation to the pivot and the indirect translation between the two other languages. This experiment is related to the setting of Nakashole and Flauger (2017), with the difference that, being in an unsupervised case, we do not have prior information about a good pivot language for a given pair, hence, the choice of a fixed pivot across the experiments. The choice of English for pivot is motivated by the fact that it is commonly used as a referential for publicly available aligned vectors.

Table 1 compares HUG with bilingual alignment (Bil.) on both direct and indirect translations. Note that, in this setting, there is no difference between MST, Star or any variation of HUG. For reference, we also report the performance obtained with a direct bilingual alignment between the 2 languages.

Overall, our approach improves indirect translation by 4.2% in average, while the overall impact on the direct translation is marginal, i.e., -0.3%. HUG significantly reduces the drop of performance of indirect alignment from 3.3% to 0.9% when compared to direct alignment. The effect seems to be more important when one of the language is poorly aligned to the pivot. Interestingly, in some cases, using a pivot can improve the performance of direct translation, typically when direct alignment is not very good. Finally, when the two languages are already well aligned with English (e.g., nl-de or it-pt), they tend to also have good unconstrained indirect translation. For these pairs, the impact of our approach is marginal, with an improvement of less than a percent.

5.2 Multilingual alignment

In this section, we evaluate the quality of joint multilingual alignment on 11 languages. We look at the impact on direct and indirect alignments. We consider two versions of HUG: HUG-r, where the tree is drawn at random at each iteration and HUG-s, where the tree is built with a MST on the distance between languages at each epoch.

Direct word translation. Table 2 shows the accuracy on direct word translations, i.e., translation to the pivot language, English. We average accuracy from and to English for each language. This setting is unfair to the MST baseline, since it does not have the same set of direct translations. For the bilingual supervised and unsupervised alignments, we learn a model per pair. In terms of baselines, our implementation of Wasserstein-Procrustes (W-Proc.*), initialized by optimizing Gromov-Wasserstein, outperforms the one introduced in Grave et al. (2018) (W-Proc.) by 2.8%.

	en-fr	en-es	en-it	en-pt	en-de	en-pl	en-ru	en-da	en-nl	en-cs	Avg.
supervised, unconstrained											
Proc.	75.9	77.1	72.0	72.8	68.2	56.9	52.2	50.6	64.9	51.7	64.2
RCSLS	80.5	83.0	-	-	73.7	-	58.7	-	-	-	-
unsupervised, unconstrained											
Adv.	66.2	70.5	-	-	61.4	-	35.3	-	-	-	
ICP	74.9	76.0	70.8	-	**67.0**	-	42.6	-	-	-	-
W-Proc.	73.5	**76.4**	68.5	68.5	64.5	45.1	40.6	46.3	62.2	41.8	58.7
W-Proc.*	**76.6**	72.1	**71.4**	72.4	66.4	**53.0**	42.8	49.0	**64.3**	46.3	61.5
unsupervised, constrained											
MST	65.3	69.6	60.7	63.8	64.7	44.6	39.3	42.9	59.1	39.5	54.8
Star	71.2	72.9	68.1	69.1	63.6	49.8	43.9	47.5	60.3	45.6	59.2
HUG-r	76.1	72.9	71.0	**72.6**	66.8	52.6	44.9	49.2	64.1	**48.1**	**61.8**
HUG-s	73.6	75.6	69.5	71.1	65.5	50.3	**45.4**	**49.4**	62.5	46.9	61.0

Table 2: Accuracy with a NN criterion on direct bilingual alignment between our approach and different bilingual unsupervised approaches as well as a supervised approaches, orthogonal Procrustes and RCSLS (Joulin et al., 2018). All the accuracies are the average of both directions (source-target and target-source). For fair comparison, we report Adv. (Conneau et al., 2017), ICP (Hoshen and Wolf, 2018) and W-Proc. (Grave et al., 2018) without the refinement step. W-Proc.* is our implementation of W-Proc. with a Gomorov-Wasserstein initialization.

On direct translations, most constrained multilingual alignments, but HUG-r, perform slightly worse than their bilingual counterparts. This is not surprising since bilingual alignment methods focus on this task, while multilingual alignment methods focus on improving indirect translation. The Star model performance drops significantly, which can be attributed to the difficulty of learning a non-convex function with $O(N^2)$ terms. This is confirmed by the performance of HUG-r ($+2.6\%$ compared to Star) where the N^2 constraints are simply replaced by a sampling of N. The gain of HUG-r over its bilingual counterpart (W-Proc.*) is too marginal ($+0.3\%$) to draw any conclusion.

	Latin	Germanic	Slavic	Latin-Germanic	Latin-Slavic	Germanic-Slavic	All
Bil.	73.4	52.8	42.5	50.7	42.5	37.7	48.5
MST	75.0	50.0	49.3	46.4	41.9	35.1	47.2
Star	74.6	52.8	47.6	50.5	44.0	39.7	49.8
HUG-r	74.6	53.5	46.8	51.4	**44.4**	40.1	50.2
HUG-s	**76.3**	**55.4**	**49.3**	**51.7**	44.0	**40.2**	**50.7**

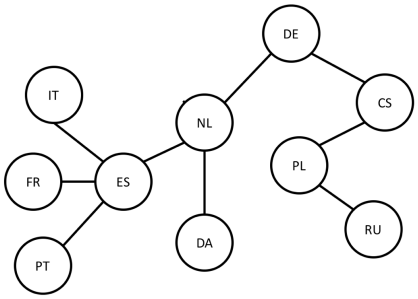
Table 3: Accuracy with a NN criterion on indirect translations averaged among and across language families. The bilingual baseline ("Bil.") is the Procrustes-Wasserstein of Grave et al. (2018) with a Gromov-Wasserstein initialization.

Indirect word translation. Table 3 shows the performance on indirect word translation with English as a pivot language. We consider averaged accuracies among and across language families, i.e. Latin, Germanic and Slavic. For the MST approach, we compute the translation by following the unique path in the tree. This can be a direct translation and it does not necessarily involve English. As expected, all of the constrained approaches improve over the bilingual baseline, by 1 to 2%. Among them, HUG-s has the best performance with an average improvement of $+2.2\%$. Note that this improvement is averaged over 45 pairs of languages, while the drop of -0.5% in direct translations is averaged over 10 pairs of languages, leading to an overall improvement of $+1.7\%$ over all 55 pairs. Similarly, the overall improvement for HUG-r is $+1.5\%$.

Interestingly, HUG-r and HUG-s improve translation for every pair of language families. The biggest improvement comes from translation among Slavic languages ($+6.8\%$). This is not surprising since it is the language family that is the most distant from the pivot, i.e., English. Similarly,

it is not surprising that the smallest improvement is between Latin and Germanic languages ($+1\%$), since English is a natural pivot between them.

Language tree. As a by-product, our approach discovers a language tree, as shown in the figure on the right. We remove English since it is not used directly. Three clusters appear: the Latin, Germanic and Slavic families chained as Latin-Germanic-Slavic. However, the pairs in the connections are harder to interpret, like Dutch-Spanish.



6 CONCLUSION

This paper introduces an unsupervised multilingual alignment method that maps every language into a common space while minimizing the impact on indirect word translation. We show that simply adding circular constraints on the mapped words significantly reduces the drop of performance in pivot word translation, and discuss several implementations of this idea. Despite its simplicity, we observe a consistent gain over bilingual translation, even sometimes on direct word translation. However, our current approach is relatively hard to scale and how to jointly learn alignment over hundreds of languages remains an open question.

REFERENCES

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, 2016.

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)*, volume 1, pages 451–462, 2017.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint*, arXiv:1607.04606, 2016. URL `https://arxiv.org/abs/1607.04606`.

Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1818–1827, 2016.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, pages 22922300, 2013.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41 (6):391–407, 1990.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 285339, 1991.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

John C Gower, Garmt B Dijksterhuis, et al. *Procrustes problems*, volume 30. Oxford University Press on Demand, 2004.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. *CoRR*, abs/1805.11222, 2018. URL `http://arxiv.org/abs/1805.11222`.

Mikhail Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. Springer Science & Business Media, 2007.

Yedid Hoshen and Lior Wolf. An iterative closest point method for unsupervised word translation. *arXiv preprint arXiv:1801.06126*, 2018.

Gary B Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Alexander Lorbert and Peter J Ramadge. Kernel hyperalignment. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2012.

Facundo Mémoli. On the use of gromov-hausdorff distances for shape comparison. 2007.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

T. Mikolov, Q. V. Le, and I Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint*, arXiv:1309.4168v, 2013a. URL `https://arxiv.org/abs/1309.4168`.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

Ndapandula Nakashole and Raphael Flauger. Knowledge distillation for bilingual dictionary induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2497–2506, 2017.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *Proceedings of ICML*, 2016.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. Technical report, 2017.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.

P. H. Schnemann. A generalized solution of the orthogonal procrustes problem. *Psychome- trika*, 31(1):110, 1966.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):72, 2016.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, 2015a.

D. Xing, C.and Wang, C. Liu, and Y Lin. Normalized word embedding and orthogonal transform for bilingual word translation. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 10061011, 2015b.

M. Zhang, Y. Liu, H. Luan, and M. Sun. Earth movers distance minimization for unsupervised bilingual lexicon induction. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 19341945, 2017a.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1959–1970, 2017b.