Gromov-Wasserstein Alignment of Word Embedding Spaces

David Alvarez-Melis
CSAIL, MIT
dalvmel@mit.edu

Tommi S. Jaakkola
CSAIL, MIT
tommi@mit.edu

Abstract

Cross-lingual or cross-domain correspondences play key roles in tasks ranging from machine translation to transfer learning. Recently, purely unsupervised methods operating on monolingual embeddings have become effective alignment tools. Current state-of-the-art methods, however, involve multiple steps, including heuristic post-hoc refinement strategies. In this paper, we cast the correspondence problem directly as an optimal transport (OT) problem, building on the idea that word embeddings arise from metric recovery algorithms. Indeed, we exploit the *Gromov-Wasserstein* distance that measures how similarities between pairs of words relate across languages. We show that our OT objective can be estimated efficiently, requires little or no tuning, and results in performance comparable with the state-of-the-art in various unsupervised word translation tasks.

1 Introduction

Many key linguistic tasks, within and across languages or domains, including machine translation, rely on learning cross-lingual correspondences between words or other semantic units. While the associated alignment problem could be solved with access to large amounts of parallel data, broader applicability relies on the ability to do so with largely mono-lingual data, from Part-of-Speech (POS) tagging (Zhang et al., 2016), dependency parsing (Guo et al., 2015), to machine translation (Lample et al., 2018). The key subtask of bilingual lexical induction, for example, while long standing as a problem (Fung, 1995; Rapp, 1995, 1999), has been actively pursued recently (Artetxe et al., 2016; Zhang et al., 2017a; Conneau et al., 2018).

Current methods for learning cross-domain correspondences at the word level rely on distributed representations of words, building on the observation that mono-lingual word embeddings exhibit similar geometric properties across languages Mikolov et al. (2013). While most early work assumed some, albeit minimal, amount of parallel data (Mikolov et al., 2013; Dinu et al., 2014; Zhang et al., 2016), recently fully-unsupervised methods have been shown to perform on par with their supervised counterparts (Conneau et al., 2018; Artetxe et al., 2018). While successful, the mappings arise from multiple steps of processing, requiring either careful initial guesses or post-mapping refinements, including mitigating the effect of frequent words on neighborhoods. The associated adversarial training schemes can also be challenging to tune properly (Artetxe et al., 2018).

In this paper, we propose a direct optimization approach to solving correspondences based on recent generalizations of optimal transport (OT). OT is a general mathematical toolbox used to evaluate correspondence-based distances and establish mappings between probability distributions, including discrete distributions such as point-sets. However, the nature of mono-lingual word embeddings renders the classic formulation of OT inapplicable to our setting. Indeed, word embeddings are estimated primarily in a relational manner to the extent that the algorithms are naturally interpreted as metric recovery methods (Hashimoto et al., 2016). In such settings, previous work has sought to bypass this lack of *registration* by jointly optimizing over a matching and an orthogonal mapping (Rangarajan et al., 1997; Zhang et al., 2017b). Due to the focus on distances rather than points, we instead adopt a relational OT formulation based on the Gromov-Wasserstein distance that measures how distances between pairs of words are mapped across languages. We show that the resulting mapping admits an efficient solution and requires little or no tuning.

In summary, we make the following contributions:

- We propose the use of the Gromov-Wasserstein distance to learn correspondences between word embedding spaces in a fully-unsupervised manner, leading to a theoretically-motivated optimization problem that can be solved efficiently, robustly, in a single step, and requires no post-processing or heuristic adjustments.
- To scale up to large vocabularies we realize an extended mapping to words not part of the original optimization problem.
- We show that the proposed approach performs on par with state-of-the-art neural network based methods on benchmark word translation tasks, while requiring a fraction of the computational cost and/or hyperparameter tuning.

2 Problem Formulation

In the unsupervised bilingual lexical induction problem we consider two languages with vocabularies V_x and V_y, represented by word embeddings $X = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$ and $Y = \{\mathbf{y}^{(j)}\}_{j=1}^{m}$, respectively, where $\mathbf{x}^{(i)} \in \mathcal{X} \subset \mathbb{R}^{d_x}$ corresponds to $w_i^x \in V_x$ and $\mathbf{y}^{(j)} \in \mathcal{Y} \subset \mathbb{R}^{d_y}$ to $w_j^y \in V_y$. For simplicity, we let $m = n$ and $d_x = d_y$, although our methods carry over to the general case with little or no modifications. Our goal is to learn an alignment between these two sets of words without any parallel data, i.e., we learn to relate $\mathbf{x}^{(i)} \leftrightarrow \mathbf{y}^{(j)}$ with the implication that w_i^x translates to w_j^y.

As background, we begin by discussing the problem of learning an explicit map between embeddings in the supervised scenario. The associated training procedure will later be used for extending unsupervised alignments (Section 3.2).

2.1 Supervised Maps: Procrustes

In the supervised setting, we learn a map $T : \mathcal{X} \to \mathcal{Y}$ such that $T(\mathbf{x}^{(i)}) \approx \mathbf{y}^{(j)}$ whenever w_j^y is a translation of w_i^x. Let $\mathbf{X}$ and $\mathbf{Y}$ be the matrices whose columns are vectors $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(j)}$, respectively. Then we can find T by solving

$$\min_{T \in \mathcal{F}} \|\mathbf{X} - T(\mathbf{Y})\|_F^2 \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm $\|A\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}$. Naturally, both the difficulty of finding T and the quality of the resulting alignment depend on the choice of space $\mathcal{F}$. A classic approach constrains T to be orthonormal matrices, i.e., rotations and reflections, resulting in the orthogonal Procrustes problem

$$\min_{\mathbf{P} \in O(n)} \|\mathbf{X} - \mathbf{P}\mathbf{Y}\|_F^2 \tag{2}$$

where $O(n) = \{\mathbf{P} \in \mathbb{R}^{n \times n} \mid \mathbf{P}^\top \mathbf{P} = \mathbf{I}\}$. One key advantage of this formulation is that it has a closed-form solution in terms of a singular value decomposition (SVD), whereas for most other choices of constraint set $\mathcal{F}$ it does not. Given an SVD decomposition $\mathbf{U}\Sigma\mathbf{V}^\top$ of $\mathbf{X}\mathbf{Y}^\top$, the solution to problem (2) is $\mathbf{P}^* = \mathbf{U}\mathbf{V}^\top$ (Schönemann, 1966). Besides obvious computational advantage, constraining the mapping between spaces to be orthonormal is justified in the context of word embedding alignment because orthogonal maps preserve angles (and thus distances), which is often the only information used by downstream tasks (e.g., for nearest neighbor search) that rely on word embeddings. (Smith et al., 2017) further show that orthogonality is required for self-consistency of linear transformations between vector spaces.

Clearly, the Procrustes approach only solves the supervised version of the problem as it requires a known correspondence between the columns of $\mathbf{X}$ and $\mathbf{Y}$. Steps beyond this constraint include using small amounts of parallel data (Zhang et al., 2016) or an unsupervised technique as the initial step to generate pseudo-parallel data (Conneau et al., 2018) before solving for $\mathbf{P}$.

2.2 Unsupervised Maps: Optimal Transport

Optimal transport formalizes the problem of finding a minimum cost mapping between two point sets, viewed as discrete distributions. Specifically, we assume two empirical distributions over embeddings, e.g.,

$$\mu = \sum_{i=1}^{n} \mathbf{p}_i \delta_{\mathbf{x}^{(i)}}, \quad \nu = \sum_{j=1}^{m} \mathbf{q}_j \delta_{\mathbf{y}^{(i)}} \tag{3}$$

where $\mathbf{p}$ and $\mathbf{q}$ are vectors of probability weights associated with each point set. In our case, we usually consider uniform weights, e.g., $\mathbf{p}_i = 1/n$ and $\mathbf{q}_j = 1/m$, although if additional information were provided (such as in the form of word frequencies), those could be naturally incorporated via $\mathbf{p}$ and $\mathbf{q}$ (see discussion at the end of Section 3). We find a *transportation map* T realizing

$$\inf_{T} \left\{ \int_{\mathcal{X}} c(\mathbf{x}, T(\mathbf{x})) d\mu(\mathbf{x}) \mid T_{\#}\mu = \nu \right\}, \tag{4}$$

where the cost $c(\mathbf{x}, T(\mathbf{x}))$ is typically just $\|\mathbf{x} - T(\mathbf{x})\|$ and $T_{\#}\mu = \nu$ implies that the source points must exactly map to the targets. However, such a map need not exist in general and we instead follow a relaxed Kantorovich's formulation. In this case, the set of transportation plans is a polytope:

$$\Pi(\mathbf{p}, \mathbf{q}) = \{\Gamma \in \mathbb{R}_+^{n \times m} \mid \Gamma \mathbb{1}_n = \mathbf{p},\ \Gamma^\top \mathbb{1}_n = \mathbf{q}\}.$$

The cost function is given as a matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$, e.g., $C_{ij} = \|\mathbf{x}^{(i)} - \mathbf{y}^{(j)}\|$. The total cost incurred by Γ is $\langle \Gamma, C \rangle := \sum_{ij} \Gamma_{ij} C_{ij}$. Thus, the discrete optimal transport (DOT) problem consists of finding a plan Γ that solves

$$\min_{\Gamma \in \Pi(\mathbf{p}, \mathbf{q})} \langle \Gamma, \mathbf{C} \rangle. \tag{5}$$

Problem (5) is a linear program, and thus can be solved exactly in $O(n^3 \log n)$ with interior point methods. However, regularizing the objective leads to more efficient optimization and often better empirical results. The most common such regularization, popularized by Cuturi (2013), involves adding an entropy penalization:

$$\min_{\Gamma \in \Pi(\mathbf{p}, \mathbf{q})} \langle \Gamma, \mathbf{C} \rangle - \lambda H(\Gamma). \tag{6}$$

The solution of this strictly convex optimization problem has the form $\Gamma^* = \operatorname{diag}(\mathbf{a})\, \mathbf{K} \operatorname{diag}(\mathbf{b})$, with $\mathbf{K} = e^{-\frac{\mathbf{C}}{\lambda}}$ (element-wise), and can be obtained efficiently via the Sinkhorn-Knopp algorithm, a matrix-scaling procedure which iteratively computes:

$$\mathbf{a} \leftarrow \mathbf{p} \oslash \mathbf{K}\mathbf{b} \quad \text{and} \quad \mathbf{b} \leftarrow \mathbf{q} \oslash \mathbf{K}^\top \mathbf{a}, \tag{7}$$

where $\oslash$ denotes entry-wise division. The derivation of these updates is immediate from the form of Γ^* above, combined with the marginal constraints $\Gamma \mathbb{1}_n = \mathbf{p}$, $\Gamma^\top \mathbb{1}_n = \mathbf{q}$ (Peyré and Cuturi, 2018).

Although simple, efficient and theoretically-motivated, a direct application of discrete OT for unsupervised word translation is not appropriate. One reason is that the mono-lingual embeddings are estimated in a relative manner, leaving, e.g., an overall rotation unspecified. Such degrees of freedom can dramatically change the entries of the cost matrix $C_{ij} = \|\mathbf{x}^{(i)} - \mathbf{y}^{(j)}\|$ and the resulting transport map. One possible solution is to simultaneously learn an optimal coupling and an orthogonal transformation (Zhang et al., 2017b). The transport problem is then solved iteratively, using $C_{ij} = \|\mathbf{x}^{(i)} - \mathbf{P}\mathbf{y}^{(j)}\|$, where $\mathbf{P}$ is in turn chosen to minimize the transport cost (via Procrustes). While promising, the resulting iterative approach is sensitive to initialization, perhaps explaining why Zhang et al. (2017b) used an adversarially learned mapping as the initial step. The computational cost can also be prohibitive (Artetxe et al., 2018) though could be remedied with additional development.

We adopt a theoretically well-founded generalization of optimal transport for pairs of points (their distances), thus in line with how the embeddings are estimated in the first place. We explain the approach in detail in the next Section.

3 Transporting across unaligned spaces

In this section we introduce the Gromov-Wasserstein distance, describe an optimization algorithm for it, and discuss how to extend the approach to out-of-sample vectors.

3.1 The Gromov Wasserstein Distance

The classic optimal transport requires a distance between vectors *across* the two domains. Such a metric may not be available, for example, when the sample sets to be matched do not belong to the same metric space (e.g., different dimension). The Gromov-Wasserstein distance (Mémoli, 2011) generalizes optimal transport by comparing the metric spaces directly instead of samples across the spaces. In other words, this framework operates on distances between pairs of points calculated within each domain and measures how these distances compare to those in the other domain. Thus, it requires a weaker but easy to define notion of *distance between distances*, and operates on pairs of points, turning the problem from a linear to a quadratic one.

Formally, in its discrete version, this framework considers two measure spaces expressed in terms of within-domain similarity matrices $(\mathbf{C}, \mathbf{p})$ and $(\mathbf{C}', \mathbf{q})$ and a loss function defined between *similarity pairs*: $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, where $L(C_{ik}, C'_{jl})$ measures the discrepancy between the distances $d(\mathbf{x}^{(i)}, \mathbf{x}^{(k)})$ and $d'(\mathbf{y}^{(j)}, \mathbf{y}^{(l)})$. Typical choices for L are $L(a, b) = \frac{1}{2}(a - b)^2$ or $L(a, b) = \mathrm{KL}(a|b)$. In this framework, $L(C_{ik}, C'_{jl})$ can also be understood as the cost of "matching" i to j *and* k to l.

All the relevant values of $L(\cdot, \cdot)$ can be put in a 4-th order tensor $\mathbf{L} \in \mathbb{R}^{N_1 \times N_1 \times N_2 \times N_2}$, where $\mathbf{L}_{ijkl} = L(C_{ik}, C'_{jl})$. As before, we seek a cou-

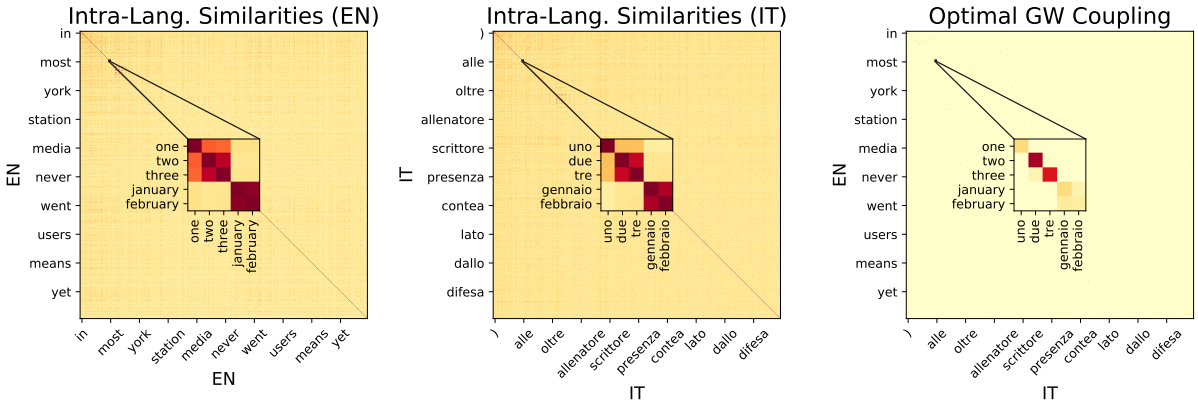


Figure 1: The Gromov-Wasserstein distance is well suited for the task of cross-lingual alignment because it relies on *relational* rather than *positional* similarities to infer correspondences across domains. Computing it requires two intra-domain similarity (or equivalently cost) matrices (**left** & **center**), and it produces an optimal coupling of source and target points with minimal discrepancy cost (**right**).

pling Γ specifying how much mass to transfer between each pair of points from the two spaces. The Gromov-Wasserstein problem is then defined as solving

$$\mathrm{GW}(\mathbf{C}, \mathbf{C}', \mathbf{p}, \mathbf{q}) = \min_{\Gamma \in \Pi(\mathbf{p},\mathbf{q})} \sum_{i,j,k,l} \mathbf{L}_{ijkl} \Gamma_{ij} \Gamma_{kl} \quad (8)$$

Compared to problem (5), this version is substantially harder since the objective is now not only non-linear, but non-convex too.[1] In addition, it requires operating on a fourth-order tensor, which would be prohibitive in most settings. Surprisingly, this problem can be optimized efficiently with first-order methods, whereby each iteration involves solving a traditional optimal transport problem (Peyré et al., 2016). Furthermore, for suitable choices of loss function L, Peyré et al. (2016) show that instead of the $O(N_1^2 N_2^2)$ complexity implied by naive fourth-order tensor product, this computation reduces to $O(N_1^2 N_2 + N1 N_2^2)$ cost. Their approach consists of solving (5) by projected gradient descent, which yields iterations that involve projecting onto $\Pi(\mathbf{p}, \mathbf{q})$ a pseudo-cost matrix of the form

$$\hat{\mathbf{C}}_\Gamma(\mathbf{C}, \mathbf{C}', \Gamma) = \mathbf{C}_{xy} - h_1(\mathbf{C}) \Gamma h_2(\mathbf{C}')^\top \quad (9)$$

where

$$\mathbf{C}_{xy} = f_1(\mathbf{C}) \mathbf{p} \mathbb{1}_m^\top + \mathbb{1}_n \mathbf{q}^\top f_2(\mathbf{C}')^\top$$

and f_1, f_2, h_2, h_2 are functions that depend on the loss L. We provide an explicit algorithm for the case $L = L_2$ at the end of this section.

[1] In fact, the discrete (Monge-type) formulation of the problem is essentially an instance of the well-known (and NP-hard) quadratic assignment problem (QAP).

Once we have solved (8), the optimal transport coupling Γ^* provides an explicit (soft) matching between source and target samples, which for the problem of interest can be interpreted as a probabilistic translation: for every pair of words $(w_{src}^{(i)}, w_{trg}^{(j)})$, Γ_{ij}^* provides a likelihood that these two words are translations of each other. This itself is enough to translate, and we show in the experiments section that Γ^* by itself, without any further post-processing, provides high-quality translations. This stands in sharp contrast to mapping-based methods, which rely on nearest-neighbor computation to infer translations, and thus become prone to hub-word effects which have to be mitigated with heuristic post-processing techniques such as Inverted Softmax (Smith et al., 2017) and Cross-Domain Similarity Scaling (CSLS) (Conneau et al., 2018). The transportation coupling Γ, being normalized *by construction*, requires no such artifacts.

The Gromov-Wasserstein problem (8) possesses various desirable theoretical properties, including the fact that for a suitable choice of the loss function it is indeed a distance:

Theorem 3.1 (Mémoli 2011)**.** *With the choice $L = L_2$, $GW^{\frac{1}{2}}$ is a distance on the space of metric measure spaces.*

Solving problem (8) therefore yields a fascinating accompanying notion: the *Gromov-Wasserstein distance between languages*, a measure of semantic discrepancy purely based on the relational characterization of their word embeddings. Owing to Theorem 3.1, such values can be

interpreted as distances, so that, e.g., the triangle inequality holds among them. In Section 4.4 we compare various languages in terms of their GW-distance.

Finally, we note that whenever word frequency counts are available, those would be used for $\mathbf{p}$ and $\mathbf{q}$. If they are not, but words are sorted according to occurrence (as they often are in popular off-the-shelf embedding formats), one can estimate rank-probabilities such as Zipf power laws, which are known to accurately model multiple languages (Piantadosi, 2014). In order to provide a fair comparison to previous work, throughout our experiments **we use uniform distributions** so as to avoid providing our method with additional information not available to others.

3.2 Scaling Up

While the pure Gromov-Wasserstein approach leads to high quality solutions, it is best suited to small-to-moderate vocabulary sizes,[2] since its optimization becomes prohibitive for very large problems. For such settings, we propose a two-step approach in which we first match a subset of the vocabulary via the optimal coupling, after which we learn an orthogonal mapping through a modified Procrustes problem. Formally, suppose we solve problem (8) for a reduced matrices $\mathbf{X}_{1:k}$ and $\mathbf{Y}_{i:k}$ consisting of the first columns k of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and let Γ^* be the optimal coupling. We seek an orthogonal matrix that best recovers the barycentric mapping implied by Γ^*. Namely, we seek to find $\mathbf{P}$ which solves:

$$\min_{\mathbf{P} \in O(n)} \|\mathbf{X}\Gamma^* - \mathbf{P}\mathbf{Y}\|_2^2 \qquad (10)$$

Just as problem (2), it is easy to show that this Procrustes-type problem has a closed form solution in terms of a singular value decomposition. Namely, the solution to (10) is $\mathbf{P}^* = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U}\Sigma\mathbf{V}^* = \mathbf{X}_{1:m}\Gamma^*\mathbf{Y}_{1:m}^\top$. After obtaining this projection, we can immediately map the rest of the embeddings via $\hat{\mathbf{y}}^{(j)} = \mathbf{P}^*\mathbf{y}^{(j)}$.

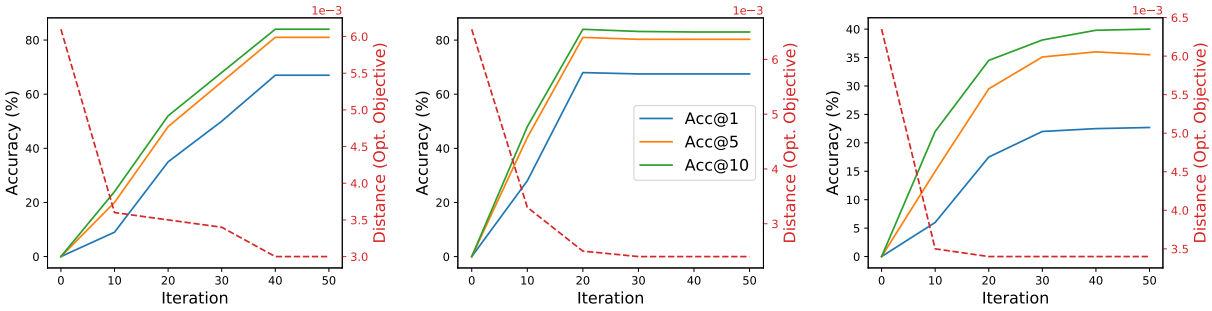
We point out that this two-step procedure resembles that of Conneau et al. (2018). Both ultimately produce an orthogonal mapping obtained by solving a Procrustes problems, but they differ in the way they produce pseudo-matches to allow for such second-step: while their approach relies on an adversarially-learned transformation, we use an explicit optimization problem.

[2] As shown in the experimental section, we are able to run problems of size in the order of $|V_s| \approx 10^5 \approx |V_t|$ on a single machine **without** relying on GPU computation.

Algorithm 1 Gromov-Wasserstein Computation for Word Embedding Alignment

Input: Source and target embeddings $\mathbf{X}$, $\mathbf{Y}$. Regularization λ. Probability vectors $\mathbf{p}, \mathbf{q}$.
// Compute intra-language similarities
$\mathbf{C}_s \leftarrow \cos(\mathbf{X}, \mathbf{X}), \quad \mathbf{C}_t \leftarrow \cos(\mathbf{Y}, \mathbf{Y})$
$\mathbf{C}_{st} \leftarrow \mathbf{C}_s^2\mathbf{p}\mathbb{1}_m^\top + \mathbb{1}_n\mathbf{q}(\mathbf{C}_t^2)^\top$
while not converged **do**
  // Compute pseudo-cost matrix (Eq. (9))
  $\hat{\mathbf{C}}_\Gamma \leftarrow \mathbf{C}_{st} - 2\mathbf{C}_s\Gamma\mathbf{C}_t^\top$
  // Sinkhorn iterations (Eq. (7))
  $\mathbf{a} \leftarrow \mathbb{1}, \quad \mathbf{K} \leftarrow \exp\{-\hat{\mathbf{C}}_\Gamma/\lambda\}$
  while not converged **do**
    $\mathbf{a} \leftarrow \mathbf{p} \oslash \mathbf{K}\mathbf{b}, \ \mathbf{b} \leftarrow \mathbf{q} \oslash \mathbf{K}^\top\mathbf{a}$
  end while
  $\Gamma \leftarrow \mathrm{diag}\,(\mathbf{a})\,\mathbf{K}\,\mathrm{diag}\,(\mathbf{b})$
end while
// Optional step: Learn explicit projection
$\mathbf{U}, \Sigma, \mathbf{V}^\top \leftarrow \mathrm{SVD}(\mathbf{X}\Gamma\mathbf{Y}^\top)$
$\mathbf{P} = \mathbf{U}\mathbf{V}^\top$
return $\Gamma, \mathbf{P}$

We end this section by discussing parameter and configuration choices. To leverage the fast algorithm of Peyré et al. (2016), we always use the L_2 distance as the loss function L between cost matrices. On the other hand, we observed throughout our experiments that the choice of cosine distance as the metric in both spaces consistently leads to better results, which agrees with common wisdom on computing distances between word embeddings. This leaves us with a single hyperparameter to control: the entropy regularization term λ. By applying any sensible normalization on the cost matrices (e.g., dividing by the mean or median value), we are able to almost entirely eliminate sensitivity to that parameter. In practice, we use a simple scheme in all experiments: we first try the same fixed value ($\lambda = 5 \times 10^{-5}$), and if the regularization proves too small (by leading to floating point errors), we instead use $\lambda = 1 \times 10^{-4}$. We never had to go beyond these two values in all our experiments. We emphasize that at no point we use train (let alone test) supervision available with many datasets—model selection is done solely in terms of the unsupervised objective. Pseudocode for the full method (with $L = L_2$ and cosine similarity) is shown here as Algorithm 1.



(a) EN →FR, 15K words, $\lambda = 5 \cdot 10^{-4}$ (b) EN →FR, 15K words, $\lambda = 10^{-4}$ (c) EN →RU, 15K words, $\lambda = 10^{-4}$

Figure 2: **Training dynamics for the Gromov-Wasserstein alignment problem**. The algorithm provably makes progress in each iteration, and the objective (red dashed line) closely follows the metric of interest (translation accuracy, not available during training). More related languages (e.g., EN →FR in 2b,2a) lead to faster optimization, while more distant pairs yield slower learning curves (EN →RU, 2c).

4 Experiments

Through this experimental evaluation we seek to: (i) understand the optimization dynamics of the proposed approach (§4.2), evaluate its performance on benchmark cross-lingual word embedding tasks (§4.3), and (iii) qualitatively investigate the notion of distance-between-languages it computes (§4.4). Rather than focusing solely on prediction accuracy, we seek to demonstrate that the proposed approach offers a fast, principled, and robust alternative to state-of-the-art multi-step methods, delivering comparable performance.

4.1 Evaluation Tasks and Methods

Datasets We evaluate our method on two standard benchmark tasks for cross-lingual embeddings. First, we consider the dataset of Conneau et al. (2018), which consists of word embeddings trained with FASTTEXT (Bojanowski et al., 2017) on Wikipedia and parallel dictionaries for 110 language pairs. Here, we focus on the language pairs for which they report results: English (EN) from/to Spanish (ES), French (FR), German (DE), Russian (RU) and simplified Chinese (ZH). We do not report results on Esperanto (EO) as dictionaries for that language were not provided with the original dataset release.

For our second set of experiments, we consider the—substantially harder[3]—dataset of (Dinu et al., 2014), which has been extensively compared against in previous work. It consists of embeddings and dictionaries in four pairs of languages; EN from/to ES, IT, DE, and FI (Finnish).

Methods To see how our fully-unsupervised method compares with methods that require (some) cross-lingual supervision, we follow (Conneau et al., 2018) and consider a simple but strong baseline consisting of solving a procrustes problem directly using the available cross-lingual embedding pairs. We refer to this method simply as PROCRUSTES. In addition, we compare against the fully-unsupervised methods of Zhang et al. (2017a), Artetxe et al. (2018) and Conneau et al. (2018).[4] As proposed by the latter, we use CSLS whenever nearest neighbor search is required, which has been shown to improve upon naive nearest-neighbor retrieval in multiple work.

4.2 Training Dynamics of G-W

As previously mentioned, our approach involves only two optimization choices, one of which is required only for very large settings. When running Algorithm 1 for the full set of embeddings is infeasible (due to memory limitations), one must decide what fraction of the embeddings to use during optimization. In our experiments, we use the largest possible size allowed by memory constraints, which was found to be $K = 20,000$ for the personal computer we used.

The other—more interesting—optimization choice involves the entropy regularization parameter λ used within the Sinkhorn iterations. Large regularization values lead to denser optimal coupling Γ^*, while less regularization leads to sparser solutions,[5] at the cost of a harder (more

[3] We discuss the difference in hardness of these two benchmark datasets in Section 4.3.

[4] Despite its relevance, we do not include the OT-based method of Zhang et al. (2017b) in the comparison because their implementation required use of proprietary software.

[5] In the limit $\lambda \to 0$, when $n = m$, the solution converges

	Supervision	Time	EN-ES		EN-FR		EN-DE		EN-IT		EN-RU	
			→	←	→	←	→	←	→	←	→	←
PROCRUSTES	5K words	3	77.6	77.2	74.9	75.9	68.4	67.7	73.9	73.8	47.2	58.2
PROCRUSTES + CSLS	5K words	3	81.2	82.3	81.2	82.2	73.6	71.9	76.3	**75.5**	51.7	63.7
(Conneau et al., 2018)	None	957	**81.7**	**83.3**	**82.3**	82.1	74.0	72.2	77.4	76.1	**52.4**	**61.4**
G-W ($\lambda = 10^{-4}$)	None	70	78.3	79.5	79.3	78.3	69.6	66.9	75.3	74.1	26.1	35.4
G-W ($\lambda = 10^{-5}$)	None	37	**81.7**	80.4	81.3	78.9	71.9	**72.8**	**78.9**	75.2	45.1	43.7

Table 1: Performance (P@1) of unsupervised and minimally-supervised methods on the dataset of Conneau et al. (2018). The time columns shows the average runtime in minutes of an instance (i.e., one language pair) of the method in this task on the same quad-core **CPU** machine.

non-convex) optimization problem.

In Figure 2 we show the training dynamics of our method when learning correspondences between word embeddings from the dataset of Conneau et al. (2018). As expected, larger values of λ lead to smoother improvements with faster runtime-per-iteration, at a price of some drop in performance. In addition, we found that computing GW distances between closer languages (such as EN and FR) leads to faster convergence than for more distant ones (such as EN and RU, in Fig. 2c).

Worth emphasizing are three desirable optimization properties that set apart the Gromov-Wasserstein distance from other unsupervised alignment approaches, particularly adversarial-training ones: (i) the objective decreases monotonically (ii) its value closely follows the true metric of interest (translation, which naturally is not available during training) and (iii) there is no risk of degradation due to *overtraining*, as is the case for adversarial-based methods trained with stochastic gradient descent (Conneau et al., 2018).

4.3 Benchmark Results

We report the results on the dataset of Conneau et al. (2018) in Table 1. The strikingly high performance of all methods on this task belies the hardness of the general problem of unsupervised cross-lingual alignment. Indeed, as pointed out by Artetxe et al. (2018), the FASTTEXT embeddings provided in this task are trained on very large and highly comparable—across languages—corpora (Wikipedia), and focuses on closely related pairs of languages. Nevertheless, we carry out experiments here to have a broad evaluation of our approach in both *easier* and *harder* settings.

Next, we present results on the more challenging dataset of (Dinu et al., 2014) in Table 2. Here, we rely on the results reported by (Artetxe et al., 2018) since by the time of writing the present work their implementation was not available yet.

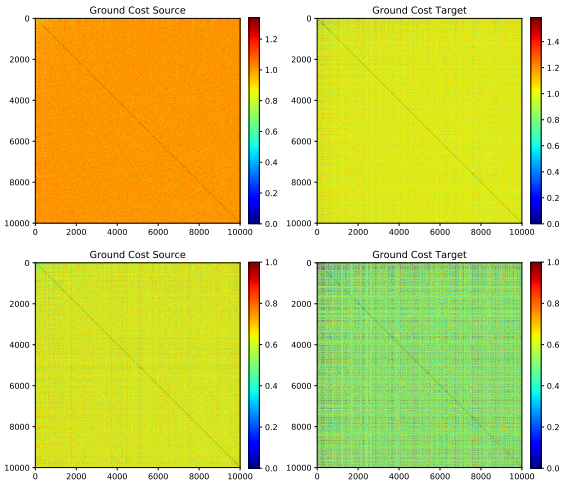


Figure 3: **Top**: Word embeddings trained on non-comparable corpora can lead to uneven distributions of pairwise distances as shown here for the EN-FI pair of (Dinu et al., 2014). **Bottom:** Normalizing the cost matrices leads to better optimization and improved performance.

Part of what makes this dataset hard is the wide discrepancy between word distance across languages, which translates into uneven distance matrices (Figure 3), and in turn leads to poor results for G-W. To account for this, previous work has relied on an initial whitening step on the embeddings. In our case, it suffices to normalize the pairwise similarity matrices to the same range to obtain substantially better results. While we have observed that careful choice of the regularization parameter λ can obviate the need for this step, we opt for the normalization approach since it allows us to optimize without having to tune λ.

We compare our method (with and without nor-

to a permutation matrix, which gives a hard-matching solution to the transportation problem (Peyré and Cuturi, 2018).

	EN-IT		EN-DE		EN-FI		EN-ES	
	P@1	Time	P@1	Time	P@1	Time	P@1	Time
(Zhang et al., 2017a)†	0	46.6	0	46.0	0.07	44.9	0.07	43.0
(Conneau et al., 2018)†	45.40	46.1	47.27	45.4	1.62	44.4	36.20	45.3
(Artetxe et al., 2018)†	48.53	8.9	**48.47**	7.3	**33.50**	12.9	**37.60**	9.1
G-W	44.4	35.2	37.83	36.7	6.8	15.6	12.5	18.4
G-W + NORMALIZE	**49.21**	36	46.5	33.2	18.3	42.1	**37.60**	38.2

Table 2: Results of unsupervised methods on the dataset of Dinu et al. (2014) with runtimes in minutes. Those marked with † are from (Artetxe et al., 2018). Note that their runtimes correspond to GPU computation, while ours are CPU-minutes, so the numbers are not directly comparable.

malization) against alternative approaches in Table 2. Note that we report the runtimes of Artetxe et al. (2018) as-is, which are obtained by running on a Titan XP GPU, while our runtimes are, as before, obtained purely by CPU computation.

4.4 Qualitative Results

As mentioned earlier, Theorem 3.1 implies that the optimal value of the Gromov-Wasserstein problem can be legitimately interpreted as a distance between languages, or more explicitly, between their word embedding spaces. This distributional notion of distance is completely determined by pairwise geometric relations between these vectors. In Figure 4 we show the values $\mathrm{GW}(\mathbf{C}_s, \mathbf{C}_t, \mathbf{p}, \mathbf{q})$ computed on the FASTTEXT word embeddings of Conneau et al. (2018) corresponding to the most frequent 2000 words in each language.

Overall, these distances conform to our intuitions: the cluster of romance languages exhibits some of the shortest distances, while classical Chinese (ZH) has the overall largest discrepancy with all other languages. But somewhat surprisingly, Russian is relatively close to the romance languages in this metric. We conjecture that this could be due to Russian's rich morphology (a trait shared by romance languages but not English). Furthermore, both Russian and Spanish are pro-drop languages (Haspelmath, 2001) and share syntactic phenomena, such as dative subjects (Moore and Perlmutter, 2000; Melis et al., 2013) and differential object marking (Bossong, 1991), which might explain why ES is closest to RU overall.

On the other hand, English appears remarkably isolated from all languages, equally distant from its germanic (DE) and romance (FR) cousins. Indeed, other aspects of the data (such as corpus size) might be underlying these observations.

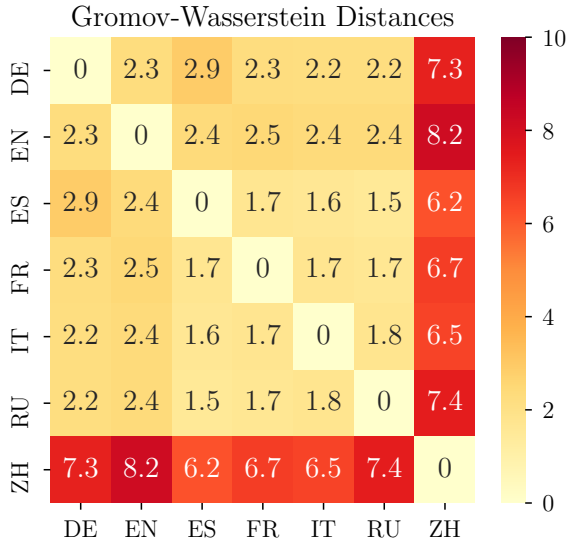


Figure 4: Pairwise language Gromov-Wasserstein distances obtained as the minimal transportation cost (8) between word embedding similarity matrices. Values scaled by 10^2 for easy visualization.

5 Related Work

Study of the problem of bilingual lexical induction goes back to Rapp (1995) and Fung (1995). While the literature on this topic is extensive, we focus here on recent fully-unsupervised and minimally-supervised approaches, and refer the reader to one of various existing surveys for a broader panorama (Upadhyay et al., 2016; Ruder et al., 2017).

Methods with coarse or limited parallel data. Most of these fall in one of two categories: methods that learn a mapping from one space to the other, e.g., as a least-squares objective (e.g., (Mikolov et al., 2013)) or via orthogonal transformations Zhang et al. (2016); Smith et al. (2017); Artetxe et al. (2016), and methods that find a com-

mon space on which to project both sets of embeddings (Faruqui and Dyer, 2014; Lu et al., 2015).

Fully Unsupervised methods. Conneau et al. (2018) and Zhang et al. (2017a) rely on adversarial training to produce an initial alignment between the spaces. The former use pseudo-matches derived from this initial alignment to solve a Procrustes (2) alignment problem. Our Gromov-Wasserstein framework can be thought of as providing an alternative to these adversarial training steps, albeit with a concise optimization formulation and producing explicit matches (via the optimal coupling) instead of depending on nearest neighbor search, as the adversarially-learnt mappings do.

Zhang et al. (2017b) also leverage optimal transport distances for the cross-lingual embedding task. However, to address the issue of non-alignment of embedding spaces, their approach follows the joint optimization of the transportation and procrustes problem as outlined in Section 2.2. This formulation makes an explicit modeling assumption (invariance to unitary transformations), and requires repeated solution of Procrustes problems during alternating minimization. Gromov-Wasserstein, on the other hand, is more flexible and makes no such assumption, since it directly deals with similarities rather than vectors. In the case where it is required, such an orthogonal mapping can be obtained by solving a single procrustes problem, as discussed in Section 3.2.

6 Discussion and future work

In this work we provided a direct optimization approach to cross-lingual word alignment. The Gromov-Wasserstein distance is well-suited for this task as it performs a relational comparison of word-vectors across languages rather than word-vectors directly. The resulting objective is concise, and can be optimized efficiently. The experimental results show that the resulting alignment framework is fast, stable and robust, yielding near state-of-the-art performance at a computational cost orders of magnitude lower than that of alternative fully unsupervised methods.

While directly solving Gromov-Wasserstein problems of reasonable size is feasible, scaling up to large vocabularies made it necessary to learn an explicit mapping via Procrustes. GPU computations or stochastic optimization could help avoid this secondary step.

Acknowledgments

The authors would like to thank the anonymous reviewers for helpful feedback. The work was partially supported by MIT-IBM grant "Adversarial learning of multimodal and structured data", and Graduate Fellowships from Hewlett Packard and CONACYT.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. 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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. 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)*, pages 789—-798. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Georg Bossong. 1991. Differential object marking in Romance and beyond. *New analyses in Romance linguistics*, pages 143–170.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *International Conference on Learning Representations*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292—-2300.

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

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Third Workshop on Very Large Corpora*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations.

In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1234–1244.

Tatsunori B Hashimoto, David Alvarez-Melis, and Tommi S Jaakkola. 2016. Word Embeddings as Metric Recovery in Semantic Spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286.

Martin Haspelmath. 2001. The European linguistic area: Standard Average European. In *Language typology and language universals: An international handbook*, volume 2, pages 1492–1510. de Gruyter.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. *International Conference on Learning Representations*.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

Chantal Melis, Marcela Flores, and A Holvoet. 2013. On the historical expansion of non-canonically marked 'subjects' in Spanish. *The diachronic Typology of Non-Canonical Subjects, Amsterdam/Philadelphia, Benjamins*, pages 163–184.

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

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *arXiv preprint arXiv:1309.4168v1*, pages 1–10.

John Moore and David M. Perlmutter. 2000. What does it take to be a dative subject? *Natural Language and Linguistic Theory*, 18(2):373–416.

Gabriel Peyré and Marco Cuturi. 2018. Computational Optimal Transport. Technical report.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. 2016. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Anand Rangarajan, Haili Chui, and Fred L Bookstein. 1997. The Softassign Procrustes Matching Algorithm. *Lecture Notes in Computer Science*, 1230:29–42.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Samuel L Smith, David H P Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *International Conference on Learning Representations*.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. 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.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, San Diego, California. Association for Computational Linguistics.