Unsupervised Hierarchy Matching with Optimal Transport over Hyperbolic Spaces

David Alvarez-Melis
Microsoft Research

Youssef Mroueh
IBM Research

Tommi Jaakkola
CSAIL, MIT

Abstract

This paper focuses on the problem of unsupervised alignment of hierarchical data such as ontologies or lexical databases. This problem arises across areas, from natural language processing to bioinformatics, and is typically solved by appeal to outside knowledge bases and label-textual similarity. In contrast, we approach the problem from a purely geometric perspective: given only a vector-space representation of the items in the two hierarchies, we seek to infer correspondences across them. Our work derives from and interweaves hyperbolic-space representations for hierarchical data, on one hand, and unsupervised word-alignment methods, on the other. We first provide a set of negative results showing how and why Euclidean methods fail in this hyperbolic setting. We then propose a novel approach based on optimal transport over hyperbolic spaces, and show that it outperforms standard embedding alignment techniques in various experiments on cross-lingual WordNet alignment and ontology matching tasks.

1 Introduction

Hierarchical structures are ubiquitous in various domains, such as natural language processing and bioinformatics. For example, structured lexical databases like WordNet (Miller, 1995) are widely used in computational linguistics as an additional resource in various downstream tasks (Moldovan and Rus, 2001; Shi and Mihalcea, 2005; Bordes et al., 2012). On the other hand, ontologies are often used to store and organize relational data. Building such datasets is expensive and requires expert knowledge, so there is great interest in methods to merge, extend and extrapolate across them. A fundamental ingredient in all of these tasks is *matching*[1] different datasets, i.e., finding correspondences between their entities. For example, the problem of ontology alignment is an active area of research, with important implications for integrating heterogeneous resources, across domains or languages (Spohr et al., 2011). We refer the reader to (Euzenat and Shvaiko, 2013) for a thorough survey on the state of this problem. On the other hand, there is a long line of work focusing on automatic WordNet construction that seeks to leverage existing large WordNets (e.g., in English) to automatically build WordNets in other low-resource languages (Lee et al., 2000; Saveski and Trajkovski, 2010; Pradet et al., 2014; Khodak et al., 2017).

Euzenat and Shvaiko (2013) recognize three dimensions to similarity in ontology matching: semantic, syntactic and external. A similar argument can be made for other types of hierarchical structures. Most current methods for aligning such types of data rely on a combination of these three, i.e., in addition to the relations between entities they exploit lexical similarity and external knowledge. For example, automatic WordNet construction methods often rely on access to machine translation systems (Pradet et al., 2014), and state-of-the-art ontology matching systems commonly assume large external knowledge bases. Unsurprisingly, these methods perform poorly when no such additional resources are available (Shvaiko and Euzenat, 2013). Thus, effective fully-unsupervised alignment of hierarchical datasets remains largely an open problem.

Our work builds upon two recent trends in machine learning to derive a new approach to this problem. On one hand, there is mounting theoretical and empirical evidence of the advantage of embedding hierarchical structures in hyperbolic (rather than Euclidean) spaces (Nickel and Kiela, 2018; Ganea et al., 2018b; De Sa et al., 2018). On the other hand, various fully unsupervised geometric approaches have recently shown remarkable success in unsupervised word translation

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

[1] Throughout this work, we interchangeably use *matching* and *alignment* to refer to this task.

(Conneau et al., 2018; Artetxe et al., 2018; Alvarez-Melis and Jaakkola, 2018; Grave et al., 2019; Alvarez-Melis et al., 2019). We seek to combine these two recent developments by extending the latter to non-Euclidean settings, and using them to find correspondences between datasets by relying solely on their geometric structure, as captured by their hyperbolic-embedded representations. The end goal is a fully unsupervised approach to the problem of hierarchy matching.

In this work, we focus on the second step of this pipeline —the *matching*— and assume the embeddings of the hierarchies are already learned and fixed. Our approach proceeds by simultaneous registration of the two manifolds and point-wise entity alignment using optimal transport distances. After introducing the building blocks of our approach, we begin our analysis with a set of negative results. We show that state-of-the-art methods for unsupervised (Euclidean) embedding alignment perform very poorly on hyperbolic embeddings, even after modifying them to account for this geometry. This failure is caused by a type of invariance —not exhibited by Euclidean embeddings— that we refer to as *branch permutation*. At a high level, this phenomenon is characterized by a lack of consistent ordering of branches in the representations of a dataset across different runs of the embedding algorithm (Fig. 1a), and is akin to the node order invariance in trees.

In response to this challenge, we further generalize our approach by learning a flexible nonlinear registration function between the spaces with a hyperbolic neural network (Ganea et al., 2018a). This nonlinear map is complex enough to register one of the hyperbolic spaces (Fig. 1b), and is learned by minimizing an optimal transport problem over hyperbolic space, which provides both a gradient signal for training and a point-wise (soft) matching between the embedded entities. The resulting method (illustrated in Fig. 3) is capable of aligning embeddings in spite of severe branch permutation, which we demonstrate with applications in WordNet translation and biological ontology matching.

In summary, we make the following contributions:

- Formulating the problem of unsupervised hierarchy matching from a geometric perspective, casting it as a correspondence problem between hyperbolic spaces
- Showing that state-of-the-art methods for unsupervised embedding alignment fail in this task, and find the cause of this to be a unique type of invariance found in popular hyperbolic embeddings
- Proposing a novel framework for Riemannian nonlinear registration based on hyperbolic neural networks, which might be of independent interest
- Empirically validating this approach with experiments on WordNet hierarchies and ontologies

Notation and Conventions Let $\mathcal{P}(\mathcal{X})$ be the set of probability distributions over a metric space $\mathcal{X}$. For a continuous map $f : \mathcal{X} \to \mathcal{Y}$ we note by $f^\sharp : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ its associated push-forward operator, i.e., for any $\mu \in \mathcal{P}(\mathcal{X})$, $\nu = f^\sharp(\mu)$ is the push-forward measure satisfying $\int_{\mathcal{Y}} h(y)\,\mathrm{d}\nu(y) = \int_{\mathcal{X}} h(f(x))\,\mathrm{d}\mu(x)\ \forall h \in \mathcal{C}(\mathcal{Y})$. The image of f is denoted as $f[\mathcal{X}] = \{f(x) \mid x \in \mathcal{X}\}$. Finally, $\mathrm{O}(n)$ and $\mathrm{SO}(n)$ are the orthogonal and special orthogonal groups of order n.

2 Related Work

Ontology Matching Ontology matching is an important problem in various bio-medical applications, e.g., to find correspondences between disease and phenotype ontologies (Algergawy et al., 2018, and references therein). Techniques in ontology matching are usually rule-based, and often rely on entity label similarity and external knowledge bases, making them unfit for unsupervised settings. Here instead we do not assume any additional information nor textual similarity.

Hyperbolic Embeddings Since their introduction (Chamberlain et al., 2017; Nickel and Kiela, 2017), research on automatic embedding of hierarchical structures in hyperbolic spaces has gained significant traction (Ganea et al., 2018b; De Sa et al., 2018; Tay et al., 2018). The main appeal of this approach is that hyperbolic geometry captures several important aspects of hierarchies and other structured data (Nickel and Kiela, 2017). We rely on these embeddings to represent the hierarchies of interest.

Unsupervised Word Embedding Alignment Word translation based on word embeddings has recently gained significant attention after the successful fully-unsupervised approach of Conneau et al. (2018). Their method finds a mapping between embedding spaces with adversarial training, after which a refinement procedure based on the Procrustes problem produces the final alignment. Various non-adversarial approaches have been proposed since, such as robust-self learning (Artetxe et al., 2018). On the other hand, Optimal transport —e.g., Wasserstein— distances have been recently shown to provide a robust and effective approach to the problem of unsupervised embedding alignment (Zhang et al., 2017; Alvarez-Melis and Jaakkola, 2018; Grave et al., 2019; Alvarez-Melis et al., 2019). For example, Alvarez-Melis et al. (2019) and Grave et al. (2019) use a hybrid optimization objective over orthogonal transformations between the spaces (i.e., an Orthogonal Procrustes problem) and Wasserstein couplings between the samples. These works consider only Euclidean settings. Alternatively, this problem can be successfully approached (Alvarez-Melis and Jaakkola, 2018) with a generalized version of optimal transport, the Gromov-Wasserstein (GW) distance

(Mémoli, 2011), which relies on comparing distances between points rather than the points themselves. The recently proposed Fused Gromov-Wasserstein distance (Vayer et al., 2018) extends this to structured domains such as graphs, but as opposed to our approach, assumes node features and knowledge of the full graph structure. While the GW distance provides a stepping stone towards alignment of more general embedding spaces, it cannot account for the type of invariances encountered in practice when operating on hyperbolic embeddings, as we will in Section 5.

Correspondence Analysis Finding correspondences between shapes is at the heart of many problems in computer graphics. One of the classic approaches to this problem is the Iterative Closest Point method (Chen and Medioni, 1992; Besl and McKay, 1992) (and its various generalizations, e.g., by Rusinkiewicz and Levoy (2001)), which alternates between finding (hard) correspondences through nearest-neighbor pairing and finding the best rigid transformation based on those correspondences (i.e., solving a Orthogonal Procrustes problem). The framework we propose can be understood as generalizing ICP in various ways: allowing for Riemannian Manifolds (beyond Euclidean spaces), going beyond rigid (orthogonal) registration and relaxing the problem by allowing for soft correspondences, which the framework of optimal transport naturally provides.

3 Hyperbolic Space Embeddings

A fundamental question when dealing with any type of symbolic data is how to represent it. As the advent of representation learning has proven, finding the right feature representation is as —and often more— important than the algorithm used on it. Naturally, the goal of such representations is to capture relevant properties of the data. For our problem, this is particularly important. Since our goal is to find correspondences between datasets based purely on their relational structure, it is crucial that the representation capture the semantics of these relations as precisely as possible.

Traditional representation learning methods embed symbolic objects into low-dimensional Euclidean spaces. These approaches have proven very successful for embedding large-scale co-occurrence statistics, like linguistic corpora for word embeddings (Mikolov et al., 2013; Pennington et al., 2014). However, recent work has shown that data for which semantics are given in the form of hierarchical structures is best represented in hyperbolic spaces, i.e., Riemannian manifolds with negative curvature (Chamberlain et al., 2017; Nickel and Kiela, 2017; Ganea et al., 2018b). Among the arguments in favor of these spaces is the fact that any tree can be embedded into finite hyperbolic spaces with arbitrary precision (Gromov, 1987). This stands in stark contrast with Euclidean spaces, for which the dependence on dimension grows exponentially. In practice, this means that very low-dimensional hyperbolic embeddings often perform on-par or above their high-dimensional Euclidean counter parts in various downstream tasks (Nickel and Kiela, 2017; Ganea et al., 2018b; Tay et al., 2018). This too is an appealing argument in our application, as we are interested in matching very large datasets, making computational efficiency crucial.

Working with hyperbolic geometry requires a model to represent it and operate on it. Recent computational approaches have mostly focused on the Poincaré Disk — or *Ball*, in higher dimensions— model. This model is defined by the manifold $\mathbb{D}^d = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$, equipped with the metric tensor $g_x^{\mathbb{D}} = \lambda_x^2 g^E$, where $\lambda_x := 1/(1 - \|x\|_2^2)$ is the *conformal factor* and g^E is the Euclidean metric tensor. With this, $(\mathbb{D}^d, g_x^{\mathbb{D}})$ has a Riemannian manifold structure, with the induced Riemannian distance given by:

$$d_{\mathbb{D}}(\mathbf{u}, \mathbf{v}) = \operatorname{arcosh}\left(1 + 2\frac{\|\mathbf{u} - \mathbf{v}\|^2}{(1-\|\mathbf{u}\|^2)(1-\|\mathbf{v}\|^2)}\right), \quad (1)$$

which yields the following norm on the Poincaré Ball:

$$\|\mathbf{u}\|_{\mathbb{D}} = d_{\mathbb{D}}(0, \mathbf{u}) = 2\operatorname{arctanh}(\|\mathbf{u}\|). \quad (2)$$

It can be seen from this expression that the magnitude of points in the Poincaré Ball tends to infinity towards its boundary. This phenomenon intuitively illustrates the tree-like structure of hyperbolic space: starting from the origin, the space becomes increasingly —in fact, exponentially more— densely packed towards the boundaries, akin to how the width of a tree grows exponentially with its depth.

Hyperbolic embedding methods find representations in the Poincaré Ball by constrained optimization (i.e., by imposing $\|\mathbf{x}\| < 1$) of a loss function that is often problem-dependent. For datasets in the form of entailment relations $\mathcal{D} = \{(u, v)\}$, where $(u, v) \in \mathcal{D}$ means that u is a subconcept of v, Nickel and Kiela (2017) propose to minimize the following soft-ranking loss:

$$\mathcal{L}(\Theta) = \sum_{(u,v)\in\mathcal{D}} \log \frac{e^{-d(\mathbf{u},\mathbf{v})}}{\sum_{\mathbf{v}'\in\mathcal{N}(u)} e^{-d(\mathbf{u},\mathbf{v}')}}, \quad (3)$$

where $\Theta = \{\mathbf{u}\}$ are the embeddings and $\mathcal{N}(u) = \{v \mid (u, v) \notin \mathcal{D}\}$ is a set of negative examples for u.

Transformations in the Poincaré Ball will play a prominent role in the development of our approach in Section 6, so we discuss them briefly here. Since the Poincaré Ball is bounded, any meaningful operation on

it must map $\mathbb{D}^d$ onto itself. Furthermore, for registration we are primarily interested in isometric transformations on the disk, i.e., we seek analogues of Euclidean vector translation, rotation and refection. In this model, translations are given by Möbius addition, defined as

$$\mathbf{u} \oplus \mathbf{v} \triangleq \frac{(1 + 2\langle \mathbf{u}, \mathbf{v}\rangle + \|\mathbf{v}\|_2^2)\mathbf{u} + (1 - \|\mathbf{u}\|_2^2)\mathbf{v}}{1 + 2\langle \mathbf{u}, \mathbf{v}\rangle + \|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2}. \quad (4)$$

This definition conforms to our intuition of translation, e.g., if the origin of the disk is translated to $\mathbf{v}$, then $\mathbf{x}$ is translated to $\mathbf{v} \oplus \mathbf{x}$. Note that this addition is neither commutative nor associative. More generally, it can be shown that all isometries in the Poincaré Ball have the form $T(\mathbf{x}) = \mathbf{P}(\mathbf{v} \oplus \mathbf{x})$, where $\mathbf{v} \in \mathbb{D}^d$ and $\mathbf{P} \in \mathrm{SO}(d)$, i.e., it is an orientation-preserving isometry in $\mathbb{R}^d$. Two other important operations are the logarithmic $\log_{\mathbf{p}}(\cdot)$ and exponential $\exp_{\mathbf{p}}(\cdot)$ maps on a Riemannian manifold, which map between the manifold and its tangent space $T_{\mathbf{p}}\mathcal{X}$ at a given point $\mathbf{p}$. For the Poincaré Ball, these maps can be expressed as

$$\exp_{\mathbf{p}}(\mathbf{u}) = \mathbf{p} \oplus \left(\tanh(\tfrac{1}{2}\lambda_{\mathbf{p}}\|\mathbf{u}\|)\tfrac{\mathbf{u}}{\|\mathbf{u}\|}\right)$$
$$\log_{\mathbf{p}}(\mathbf{v}) = \tfrac{2}{\lambda_{\mathbf{p}}} \operatorname{arctanh}\left(\|(-\mathbf{p}) \oplus \mathbf{v}\|\right) \tfrac{(-\mathbf{p})\oplus\mathbf{v}}{\|(-\mathbf{p})\oplus\mathbf{v}\|}.$$

4 Matching via Optimal Transport

4.1 Optimal Transport Distances

Optimal transport (OT) distances provide a powerful and principled approach to find correspondences across distributions, shapes, and point clouds (Villani, 2008; Peyré and Cuturi, 2018). In its usual formulation, OT considers complete and separable metric spaces $\mathcal{X}, \mathcal{Y}$, along with probability measures $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{Y})$. These can be continuous or discrete measures, the latter often used in practice as empirical approximations of the former whenever working in the finite-sample regime. Kantorovitch (1942) formulated the transportation problem as:

$$\mathrm{OT}(\alpha, \beta) \triangleq \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\pi(x, y) \quad (5)$$

where $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is a cost function (the "ground" cost), and the set of couplings $\Pi(\alpha, \beta)$ consists of joint probability distributions over the product space $\mathcal{X} \times \mathcal{Y}$ with marginals α and β, i.e.,

$$\Pi(\alpha, \beta) \triangleq \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid P_1^\sharp \pi = \alpha, P_2^\sharp \pi = \beta\}, \quad (6)$$

where $P_1(x, y) = x$ and $P_2(x, y) = y$. Whenever $\mathcal{X}, \mathcal{Y}$ are equipped with a metric d, it is natural to use it as ground cost, e.g., $c(x, y) = d(x, y)^p$ for some $p \geq 1$. In such case, $\mathrm{W}_p(\alpha, \beta) \triangleq \mathrm{OT}(\alpha, \beta)^{1/p}$ is called the p-Wasserstein distance. The case $p = 1$ is also known as the Earth Mover's Distance (Rubner et al., 2000).

In applications, the measures α and β are often unknown, and are accessible only through finite samples $\{\mathbf{x}^{(i)}\} \in \mathcal{X}, \{\mathbf{y}^{(j)}\} \in \mathcal{Y}$. In that case, these can be taken to be discrete measures $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{\mathbf{x}^{(i)}}$ and $\beta = \sum_{i=1}^m \mathbf{b}_i \delta_{\mathbf{y}^{(j)}}$, where $\mathbf{a}$, $\mathbf{b}$ are vectors in the probability simplex, and the pairwise costs can be compactly represented as an $n \times m$ matrix $\mathbf{C}$, i.e., $\mathbf{C}_{ij} = c(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})$. In this case, Equation (5) becomes a linear program. Solving this problem scales cubically on the sample sizes, which is often prohibitive in practice. Adding an entropy regularization, namely

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \triangleq \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\pi + \varepsilon \mathrm{H}(\pi \,|\, \alpha \otimes \beta) \quad (7)$$

where $\mathrm{H}(\pi \mid \alpha \otimes \beta) = \int \log(\mathrm{d}\pi / \mathrm{d}\alpha\, \mathrm{d}\beta)\, \mathrm{d}\pi$, leads to a problem that can be solved much more efficiently (Altschuler et al., 2017) and which has better sample complexity (Genevay et al., 2019) than the unregularized problem. In the discrete case, Problem (7) can be solved with the Sinkhorn algorithm (Cuturi, 2013; Peyré and Cuturi, 2018), which iteratively updates $\mathbf{u} \leftarrow \mathbf{a} \oslash \mathbf{K}\mathbf{v}$ and $\mathbf{v} \leftarrow \mathbf{b} \oslash \mathbf{K}^\top \mathbf{u}$, where $\mathbf{K} \triangleq \exp\{-\frac{1}{\varepsilon}\mathbf{C}\}$ and the division $\oslash$ and exponential are entry-wise.

4.2 Unsupervised Matching with OT

Besides providing a principled geometric approach to compare distributions, OT has the advantage of producing, as an intrinsic part of its computation, a realization of the optimal way to match the two distributions. Any feasible coupling $\pi \in \Pi(\alpha, \beta)$ in problem (5) —or (7)— can be interpreted as a "soft" (multivalued) matching between α and β. Therefore, the optimal π^* corresponds to the minimum-cost way to match them. In the case where the distributions are discrete (e.g., point clouds) π^* is a matrix of soft correspondences. Whenever OT is used with the goal of *transportation* (as opposed to just *comparison*), having guarantees on the solution of the problem takes particular importance. Obtaining such guarantees is an active area of research, and a full exposition falls beyond the scope of this work. We provide a brief summary of these in Appendix E, but refer the interested reader to the survey by Ambrosio and Gigli (2013). For our purposes, is suffices to mention that for the quadratic cost (i.e., the 2-Wasserstein distance), the optimal coupling π^* is guaranteed to exist, be unique, and correspond to a deterministic map (i.e., a "hard" matching).[2]

It is tempting to directly apply OT to unsupervised embedding alignment. But note that Problem (5) crucially requires that the metric d be congruent across the two spaces (in the Wasserstein case), or more generally, that a meaningful cost function c between them be

[2] Note that $\pi(\alpha, \beta)$ "includes" all maps $T : \mathcal{X} \to \mathcal{Y}$, which can be expressed as $\pi(\cdot, \cdot) = (Id \times T)^\sharp \alpha$.

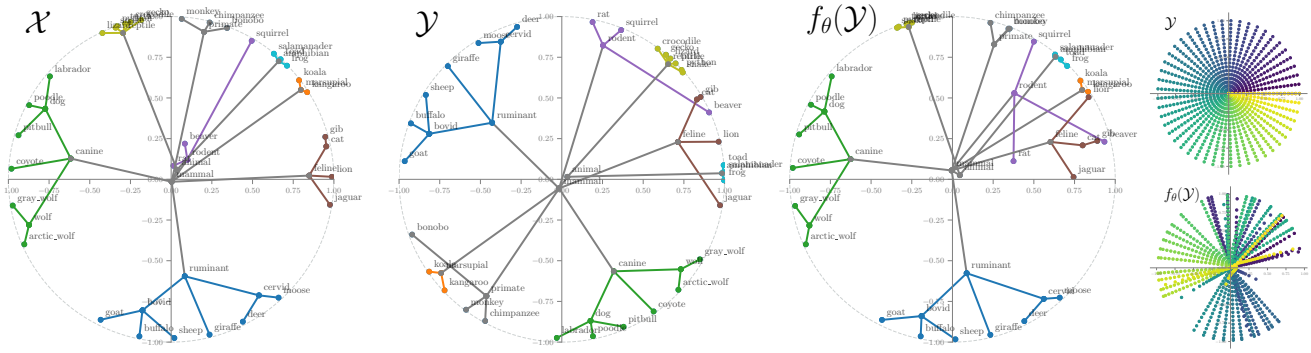

Figure 1: **Invariance and Registration in the Poincaré Disk**. Besides rotation invariance, hyperbolic embeddings can exhibit *branch permutation invariance*, characterized by changes in the relative position of branches (e.g., green and blue here) across embedding instances despite their shape being preserved. The two embeddings $\mathcal{X}$, $\mathcal{Y}$ shown here were produced by the method of Nickel and Kiela (2018) on the same simple hierarchy, using the same hyperparmeters but different random seeds. To allow for unsupervised correspondences between the embeddings to be inferred, the target space must be warped with a registration function f_θ (learnt as part of our method), which is highly non-linear, as shown by the density plots on the right.

specified. When the embedding spaces are estimated in a data-driven way, as is usually the case in machine learning, even if these spaces are compatible (e.g., have the same dimensionality) there is no guarantee that the usual metric $d(x, y)$ is meaningful. This could be, for example, because the spaces are defined up to rotations and reflections, creating a class of invariants that the ground metric does not take into account.

A natural approach to deal with this lack of registration between the two spaces is to simultaneously find a global transformation that corrects for this *and* an optimal coupling that minimizes the transportation cost between the distributions. Formally, in addition to the optimal coupling, we now also seek a mapping $f : \mathcal{Y} \to \mathcal{X}$ which realizes

$$\min_{f \in \mathcal{F}, \pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} d(x, f(y))\, \mathrm{d}\pi(x, f(y)). \qquad (8)$$

where $\mathcal{F}$ is some function class encoding the invariances.

As before, we can additionally define an entropy-regularized version of this problem too. Variations of this problem for particular cases of $(\mathcal{X}, \mathcal{Y}), d(\cdot,\cdot)$ and $\mathcal{F}$ have been proposed in various contexts, particularly for image registration (e.g., Rangarajan et al., 1997; Cohen and Guibas, 1999), and more recently, for word embedding alignment (Zhang et al., 2017; Alvarez-Melis et al., 2019; Grave et al., 2019). Virtually all these approaches instantiate $\mathcal{F}$ as the class of orthogonal transformations $\mathrm{O}(d)$, or slightly more general classes of linear mappings (Alvarez-Melis et al., 2019). In such cases, minimization with respect to f is easy to compute, as it corresponds to an Orthogonal Procrustes problem, which has a closed form solution (Gower and Dijksterhuis, 2004). Thus, Problem (8) is commonly solved by alternating minimization.

5 Hyperbolic Optimal Transport

In the previous section, we discussed how Wasserstein distances can be used to find correspondences between two embedding spaces in a fully unsupervised manner. However, all the methods we mentioned there have been applied exclusively to Euclidean settings. One might be hopeful that naive application of those approaches on hyperbolic embeddings might just work, but —unsurprisingly— it does not (cf. Table 1b). Indeed, ignoring the special geometry of these spaces leads to poor alignment. Thus, we now investigate how to adapt such a framework to non-Euclidean settings.

The first fundamental question towards this goal is whether OT extends to more general Riemannian manifolds. The answer is mostly positive. Again, limited space prohibits a nuanced discussion of this matter, but for our purposes it suffices to say that for hyperbolic spaces, under mild regularity assumptions, it can be shown that: (i) OT is well-defined (Villani, 2008), (ii) its solution is guaranteed to exist, be unique and be induced by a transport map (McCann, 2001); and (iii) this map is not guaranteed to be smooth for the usual cost $d_{\mathbb{D}}(x, y)^2$, but it is for variations of it (e.g., $-\cosh \circ d_{\mathbb{D}}$) (Lee and Li, 2012). Further details on why this is the case are provided in Appendix F. These properties justify the use of Wasserstein distances for finding correspondences in the hyperbolic setting of interest. Furthermore, Theorem F.2 provides various Riemannian cost functions with strong theoretical foundations and potential for better empirical performance.

The second step towards generalizing Problem (8) to hyperbolic spaces involves the transformation $f \in \mathcal{F}$. First, we note that using orthogonal matrices as in the Euclidean case is still valid because, as discussed in Section 3, these map the unit disk into itself. Therefore,

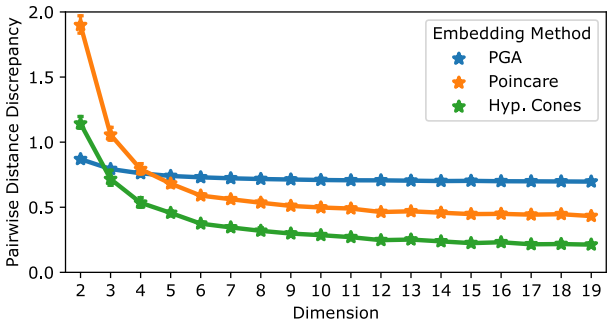

Figure 2: The branch permutation invariance occurs across hyperbolic embedding methods, even in high dimensions. Here we show the discrepancy (mean $\pm$ 1 s.d. error bar) over 5 repetitions between pairwise distance matrices of embeddings obtained with different random seeds on the same dataset (`wordnet-mammals`). Truly isometric embeddings would yield 0 discrepancy.

we can now solve a generalized (hyperbolic) version of the Orthogonal Procrustes problem as before. However, this approach performs surprisingly bad in practice too (see results for HyperOT+Orthogonal $\mathbf{P}$ in Table 1b).

To understand this failure, recall that orthogonality was a natural choice of invariance for embedding spaces that were assumed to differ at most by a rigid transformation, but be otherwise compatible. However, Poincaré embeddings exhibit another, more complex, type of invariance, which to the best of our knowledge has not been reported before. It is a *branch permutation invariance*, whereby the relative positions of branches in the hierarchy might change abruptly across different runs of the embedding algorithm, even for the exact same data and hyperparameters. This phenomenon is shown for a simple hierarchy embedded in the Poincaré Disk in Figure 1. While actual discrete trees are indeed invariant to node ordering, *a priori* it is not obvious why this property would be inherited by the embedded space obtained with objective (3), where non-ancestrally-related nodes do indeed interact (as negative pairs) in the objective. Figure 2 shows that this phenomenon occurs across hyperbolic embedding methods, such as those of Nickel and Kiela (2017) (Poincaré), Ganea et al. (2018b) (Hyperbolic Cones) and Sala et al. (2018) (Principal Geodesic Analysis), and that, although less prominent, it is still present in high dimensions.

We conjecture that the cause of this invariance is the use of negative sampling in the embedding objectives, which prioritizes distance preservation between entities that are ancestrally related in the hierarchy, at the cost of downweighting distances between unrelated entities. A formal explanation of this phenomenon is left for future work. Here, instead, we develop a framework to account and correct these invariances while simultaneously aligning the two embeddings.

6 A Framework for Correspondence across Hyperbolic Spaces

The failure of the baseline Euclidean alignment methods (and their hyperbolic versions), combined with the underlying branch permutability invariance responsible for it, make it clear that the space of registration transformations $\mathcal{F}$ in Problem (8) has to be generalized not only beyond orthogonality but beyond linearity too.

Ideally we would search among all continuous mappings between $\mathcal{Y}$ and $\mathcal{X}$, i.e, taking $\mathcal{F} = \{f : \mathcal{Y} \to \mathcal{X} \mid f \in \mathcal{C}(\mathcal{Y})\}$. To make this search computationally tractable, we can instead approximate this function class with deep neural networks f_θ parametrized by $\theta \in \Theta$. While an alternating minimization approach is still possible, solving for θ to completion in each iteration is undesirable. Instead, we reverse the order of optimization and rewrite our objective $\min_{f_\theta \in \mathcal{F}} \mathrm{W}_\varepsilon(\alpha, f_\theta^\sharp \beta)$ as

$$\min_{f_\theta, \pi} \int_{\mathcal{X}\times\mathcal{Y}} d(x, f_\theta(y))\, \mathrm{d}\pi(x, f_\theta(y)) + \varepsilon \mathrm{H}(\pi \,|\, \alpha \otimes \beta).$$

Since $\mathrm{W}_\varepsilon(\alpha, f_\theta^\sharp \beta)$ is differentiable with respect to θ, we can use gradient-based methods to optimize it. Wasserstein distances have been used as loss functions before, particularly in deep generative modeling (Arjovsky et al., 2017; Genevay et al., 2018; Salimans et al., 2018). When used as a loss function, the entropy-regularized version (Eq. (7)) has the undesirable property that $\mathrm{W}_\varepsilon(\alpha, \alpha) \neq 0$, in addition to having biased sample gradients (Bellemare et al., 2017). Following Genevay et al. (2018), we instead consider the *Sinkhorn Divergence*:

$$\mathrm{SD}_\varepsilon(\alpha, \beta) \triangleq \mathrm{W}_\varepsilon(\alpha, \beta) - {}^1\!/\!_2\big(\mathrm{W}_\varepsilon(\alpha, \alpha) + \mathrm{W}_\varepsilon(\beta, \beta)\big). \quad (9)$$

Besides being a true divergence and providing unbiased gradients, this function is convex, smooth and positive-definite (Feydy et al., 2019), and its sample complexity is well characterized (Genevay et al., 2019), all of which make it an appealing loss function. Using this divergence in place of the Wasserstein distance above yields our final objective:

$$\min_{\theta : f_\theta[\mathbb{D}^d] \subseteq \mathbb{D}^d} \mathrm{SD}_\varepsilon(\alpha, f_\theta^\sharp \beta). \quad (10)$$

The last remaining challenge is that we need to construct a class of neural networks that parametrizes $\mathcal{F} := \{f_\theta \mid f_\theta(\mathbb{D}^d) \subseteq \mathbb{D}^d\}$, i.e., functions that map $\mathbb{D}^d$ onto itself. In recent work, Ganea et al. (2018b) propose a class of hyperbolic neural networks that do precisely this. As they point out, the basic operations in hyperbolic space that we introduced in Section 3 suffice to define analogues of various differentiable building blocks of traditional neural networks. For example, a hyperbolic linear layer can be defined as

$$f_{\mathrm{HypL}}(\mathbf{x}; \mathbf{W}, \mathbf{b}) \triangleq (\mathbf{W} \otimes \mathbf{x}) \oplus \mathbf{b} = \exp_0(\mathbf{W} \log_0(\mathbf{x})) \oplus \mathbf{b}.$$

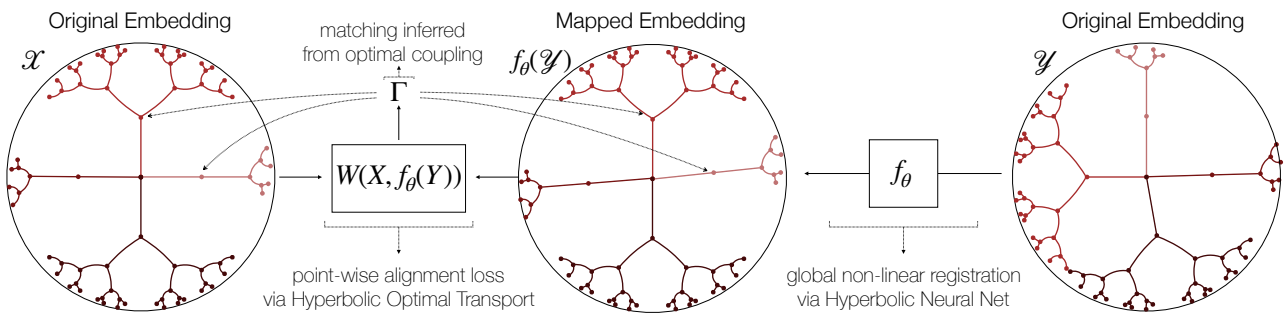

Figure 3: Schematic representation of the proposed approach. A deep network f_θ globally registers the two hyperbolic embedding spaces ($\mathcal{X}$ and $\mathcal{Y}$) by correcting for non-linear branch permutations, so that source and mapped target points can be aligned using a hyperbolic variant of Wasserstein distance. Training is done end-to-end in a fully unsupervised way – no prior known correspondences between the hierarchies are assumed.

Analogously, applying a nonlinearity $\sigma(\cdot)$ in the hyperbolic sense can be defined as $\sigma_{\mathbf{D}}(\mathbf{x}) \triangleq \exp_0(\sigma \log_0(\mathbf{x}))$. Here, we also consider *Möbius Transformation layers*, $f_{\text{Möbius}}(\mathbf{x}) = \mathbf{P}(\mathbf{v} \oplus \mathbf{x})$, with $\mathbf{P} \in \text{SO}(d)$ and $\mathbf{v} \in \mathbb{D}^d$. With these building blocks, we can parametrize highly nonlinear maps $f_\theta : \mathbb{D}^n \to \mathbb{D}^n$ as a sequence of such hyperbolic layers, i.e., $\mathbf{h}^{(i)} = \sigma_{\mathbb{D}}(\mathbf{W}^{(i)} \otimes \mathbf{h}^{(i-1)} \oplus \mathbf{b}^{(i)})$. Note that the hyperbolic linear —but *not* the Möbius— layers allow for intermediate vectors $\mathbf{h}^{(i)}$ that are of different size as the input and output, i.e., we can use rectangular weight matrices $\mathbf{W}$ to map intermediate states to Poincaré balls of different dimensionality.

The overall approach is summarized in Figure 3.

Optimization Evaluation of the loss function in (10) is itself an optimization problem, i.e., solving instances of regularized optimal transport. We backpropagate through this objective as proposed by Genevay et al. (2018), using the `geomloss` toolbox for efficiency. For the outer-level optimization, we rely on Riemannian gradient descent (Zhang et al., 2016; Wilson and Leimeister, 2018). We found that the adaptive methods of Bécigneul and Ganea (2019) worked best, particularly Radam. Note that for the HyperLinear layers only the bias term is constrained (on the Poincaré Ball), while for our Möbius layers the weight matrix is also constrained (in the Stiefel manifold), hence for these we optimize over the product of these two manifolds. Additional details are provided in the Appendix.

Avoiding poor local minima The loss function (10) is highly non-convex with respect to θ, a consequence of both the objective itself and the nature of hyperbolic neural networks (Ganea et al., 2018b). As a result, we found that initialization plays a crucial role in this problem, since it is very hard to overcome a poor initial local minimum. Even suitable layer-wise random initialization of weights and biases proved futile. As a solution, we experimented with three pre-training initialization schemes, that roughly ensure (in different ways) that f does not initially "collapse" the space $\mathcal{Y}$ (details provided in Appendix A). In addition, we use an annealing scheme on the entropy-regularization parameter ε (Kosowsky and Yuille, 1994; Alvarez-Melis et al., 2019). Starting from an aggressive regularization (large ε_0), we gradually decrease it with a fixed decay rate $\varepsilon_t = \xi \cdot \varepsilon_{t-1}$. We use $\xi = 0.99$ in all experiments.

7 Experiments

Datasets For our first set of experiments, we extract subsets of WordNet (Miller, 1995) in five languages. For this, we consider only nouns and compute their transitive closure according to hypernym relations. Then, for each collection we generate embeddings in the Poincaré Ball of dimension 10 using the method of Nickel and Kiela (2018) with default parameters. We release the multi-lingual WordNet dataset along with our codebase. In Section 7.2, we perform synthetic experiments on the `cs-phd` network dataset (Rossi and Ahmed, 2015) again embedded with same algorithm. In addition, we consider two subtasks of the OAEI 2018 ontology matching challenge (Algergawy et al., 2018): Anatomy, which consists of two ontologies; and bio-div, consisting of four. Additional details on all the datasets are provided in Table 3 in the Appendix.

Methods We first compare ablated versions of our Hyperbolic-OT model (Table 4 in the Appendix), and then we compare against three state-of-the-art unsupervised word embedding alignment models: Muse (Conneau et al., 2018), Self-Learn (Artetxe et al., 2018) and InvarOt (Alvarez-Melis et al., 2019), using their recommended configurations and settings.

Metrics All the baseline methods return transformed embeddings. Using these, we retrieve nearest neighbors and report precision-at-k, i.e., P@k = α if the true match is within the top k retrieved candidate matches for α percent of the test examples.

	P@1	P@10
FULL MODEL	22.2	88.8
small	8.7	38.0
euclidean	3.1	13.0
elu→relu	6.9	37.6
RADAM→RSGD	14.7	69.5
MÖBIUS layers	11.9	54.3
cost: $-\cosh \circ d$	16.6	70.2
no pretrain	0.1	0.2

(a) Ablation on EN→EN.

		EN-ES		EN-IT		EN-FR		EN-CA	
	T(m)	→	←	→	←	→	←	→	←
Baseline Euclidean									
MUSE	42	0.60	1.20	0.06	0.45	0.06	2.22	0.09	0.87
SELF-LEARN	3	0.31	1.40	0.46	0.46	0.55	0.40	0.25	0.34
INVAR-OT	4	0.25	0.25	2.15	0.60	0.38	2.14	0.51	7.65
Hyperbolic Methods									
HYP-OT + Orth. P	13	4.51	5.29	11.2	0.47	5.32	4.68	7.21	5.55
HYP-OT + NN f_θ	21	43.9	56.8	48.0	60.1	54.3	57.5	38.4	57.4

(b) P@10 for pair-wise language matching. T(m): runtime in minutes.

Table 1: Results on the mutlilingual wordnet matching task.

7.1 Multilingual WordNet Alignment

We first perform an ablation study on the various components of our model in a controlled setting, where the correspondences between the two datasets are perfect and unambiguous. For this, we embed the same hierarchy (the EN part of our WordNet dataset) twice, using the same algorithm with the same hyperparameters, but different random seeds. We then evaluate the extent to which our method can recover the correspondences. Starting from our FULL MODEL, we remove and/or replace various components and re-evaluate performance. The exact configuration of the ablated models is provided in Appendix D. The results in Table 1a suggest that the most crucial components are the use of the appropriate Poincaré metric and the pretraining step. Next, we move on to the real task of interest: matching WordNet embeddings across different languages. Naturally, in this case there might not be perfect correspondences across the entities in different languages. As before, we report Precision@10 and compare against baseline models in Table 1b.

7.2 Noise Sensitivity

We next analyze the effect of domain discrepancy on the matching quality. Using our method to match a noise-less and noisy versions of the `csphd` dataset (details in the Appendix), we observe that accuracy degrades rapidly with noise, although —as expected— less so for higher-dimensional embeddings (Fig. 4).

7.3 Ontology Matching

Finally, we test our method on the OAEI tasks. The results (Table 2) show that our method again decidedly outperforms the baseline Euclidean methods, but now the overall performance of all methods is remarkably lower, which suggests the domains' geometry is less coherent (partly because of their vastly different sizes) and/or the correspondences between them more noisy.

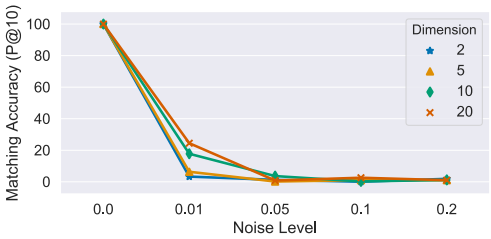

Figure 4: Noise sensitivity.

	Anatomy		Biodiv			
	H→M	M→H	F→P	P→F	E→S	S→E
MUSE	0.12	0.00	3.23	0.00	0.00	0.00
SELF-LEARN	0.00	0.00	4.00	0.00	0.01	0.02
HYP-OT	7.89	4.49	16.67	8.73	6.25	9.66

Table 2: Results on the OAEI ontology matching tasks.

8 Discussion and Extensions

The framework for hierarchical structure matching proposed here admits various extensions, some of them immediate. We focused on the particular case of the Poincaré Ball, but since most of the components of our approach —optimization, registration, optimal transport— generalize to other Riemannian manifolds, our framework would too. As long as optimizing over a given manifold is tractable, our framework would enable computing correspondences across instances of it. On the other hand, we purposely adopted the challenging setting where no additional information is assumed. This setting is relevant both for extreme practical cases and to stress-test the limits of unsupervised learning in this context. However, our method would likely benefit from incorporating any additional available information as state-of-the-art methods for ontology matching do. In our framework, this information could for example be injected intro the transport cost objective.

Acknowledgements

The work was partially supported by the MIT-IBM project on adversarial learning of multi-modal and structured data.

References

Algergawy, A. et al. (2018). "Results of the Ontology Alignment Evaluation Initiative 2018". In: CEUR-WS: Workshop proceedings.

Altschuler, J., J. Weed, and P. Rigollet (2017). "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems*, pp. 1961–1971.

Alvarez-Melis, D. and T. S. Jaakkola (2018). "Gromov-Wasserstein Alignment of Word Embedding Spaces". In: *Empirical Methods in Natural Language Processing*.

Alvarez-Melis, D., S. Jegelka, and T. S. Jaakkola (2019). "Towards Optimal Transport with Global Invariances". In: *Artificial Intelligence and Statistics*.

Ambrosio, L. and N. Gigli (2013). "A User's Guide to Optimal Transport". In: *Modelling and Optimisation of Flows on Networks: Cetraro, Italy 2009, Editors: Benedetto Piccoli, Michel Rascle*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–155.

Arjovsky, M., S. Chintala, and L. Bottou (2017). "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. PMLR, pp. 214–223.

Artetxe, M., G. Labaka, and E. 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)*. Association for Computational Linguistics, pp. 789–798.

Bécigneul, G. and O.-E. Ganea (2019). "Riemannian Adaptive Optimization Methods". In: *International Conference on Learning Representations*.

Bellemare, M. G., I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos (2017). "The cramer distance as a solution to biased wasserstein gradients". In: *arXiv preprint arXiv:1705.10743*.

Besl, P. J. and N. D. McKay (1992). "A Method for Registration of 3-D Shapes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bordes, A., X. Glorot, J. Weston, and Y. Bengio (2012). "Joint learning of words and meaning representations for open-text semantic parsing". In: *Artificial Intelligence and Statistics*, pp. 127–135.

Brenier, Y. (1987). "Décomposition polaire et réarrangement monotone des champs de vecteurs". In: *CR Acad. Sci. Paris Sér. I Math* 305.19, pp. 805–808.

– (1991). "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on Pure and Applied Mathematics*.

Bunne, C., D. Alvarez-Melis, S. Jegelka, and A. Krause (2019). "Learning Generative Models Across Incomparable Spaces". In: *International Conference on Machine Learning*.

Caffarelli, L. A. (1992a). "Boundary regularity of maps with convex potentials". In: *Communications on Pure and Applied Mathematics*.

Caffarelli, L. A. (1992b). "The regularity of mappings with a convex potential". In: *Journal of the American Mathematical Society* 5.1, pp. 99–104.

Chamberlain, B. P., J. Clough, and M. P. Deisenroth (2017). "Neural embeddings of graphs in hyperbolic space". In: *arXiv preprint arXiv:1705.10359*.

Chen, Y. and G. Medioni (1992). "Object modelling by registration of multiple range images". In: *Image and Vision Computing*.

Cohen, S. and L. Guibas (1999). "The Earth Mover's Distance under transformation sets". In: *Proceedings of the 7th IEEE International Conference on Computer Vision*.

Conneau, A., G. Lample, M. Ranzato, L. Denoyer, and H. Jégou (2018). "Word Translation Without Parallel Data". In: *International Conference on Learning Representations*, pp. 1–13.

Cuturi, M. (2013). "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*, pp. 2292–2300.

De Sa, C., A. Gu, C. Ré, and F. Sala (2018). "Representation tradeoffs for hyperbolic embeddings". In: *arXiv preprint arXiv:1804.03329*.

Euzenat, J. and P. Shvaiko (2013). *Ontology matching*. 2nd. Heidelberg (DE): Springer-Verlag.

Feydy, J., T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve, and G. Peyré (2019). "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". In: *Artificial Intelligence and Statistics*. Vol. 89. PMLR, pp. 2681–2690.

Ganea, O., G. Bécigneul, and T. Hofmann (2018a). "Hyperbolic neural networks". In: *Advances in Neural Information Processing Systems*, pp. 5345–5355.

Ganea, O.-E., G. Bécigneul, and T. Hofmann (2018b). "Hyperbolic Entailment Cones for Learning Hierarchical Embeddings". In: *International Conference on Machine Learning*.

Genevay, A., G. Peyre, and M. Cuturi (2018). "Learning Generative Models with Sinkhorn Divergences". In: *International Conference on Artificial Intelligence and Statistics*. Vol. 84. PMLR, pp. 1608–1617.

Genevay, A., L. Chizat, F. Bach, M. Cuturi, and G. Peyré (2019). "Sample complexity of Sinkhorn divergences". In: *Artificial Intelligence and Statistics*.

Gower, J. C. and G. B. Dijksterhuis (2004). *Procrustes problems*. Vol. 30. Oxford University Press on Demand.

Grave, E, A Joulin, and Q Berthet (May 2019). "Unsupervised Alignment of Embeddings with Wasserstein Procrustes". In: *Artificial Intelligence and Statistics*.

Gromov, M. (1987). "Hyperbolic groups". In: *Essays in group theory*. Springer, pp. 75–263.

Kantorovitch, L. (1942). "On the Translocation of Masses". In: *Dokl. Akad. Nauk SSSR* 37.7-8, pp. 227–229.

Khodak, M., A. Risteski, C. Fellbaum, and S. Arora (2017). "Automated wordnet construction using word embeddings". In: *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 12–23.

Kosowsky, J. J. and A. L. Yuille (Mar. 1994). "The Invisible Hand Algorithm: Solving the Assignment Problem with Statistical Physics". In: *Neural Netw.* 7.3, pp. 477–490.

Lee, C., G. Leer, and S. JungYun (2000). "Automatic WordNet mapping using word sense disambiguation". In: *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Lee, P and J Li (2012). "New Examples Satisfying Ma-Trudinger-Wang Conditions". In: *SIAM Journal on Mathematical Analysis* 44.1, pp. 61–73.

Li, J. (2009). "Smooth Optimal Transport on Hyperbolic Space". PhD thesis. University of Toronto.

Loeper, G. (2009). "On the regularity of solutions of optimal transportation problems". In: *Acta Mathematica*.

Ma, X. N., N. S. Trudinger, and X. J. Wang (2005). "Regularity of potential functions of the optimal transportation problem". In: *Archive for Rational Mechanics and Analysis*.

McCann, R. J. (2001). "Polar factorization of maps on Riemannian manifolds". In: *Geometric and Functional Analysis*.

Mémoli, F. (2011). "Gromov–Wasserstein distances and the metric approach to object matching". In: *Foundations of computational mathematics* 11.4, pp. 417–487.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *NIPS*, pp. 1–9.

Miller, G. A. (1995). "WordNet: a lexical database for English". In: *Communications of the ACM*.

Moldovan, D. and V. Rus (2001). "Logic form transformation of wordnet and its applicability to question answering". In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.

Nickel, M. and D. Kiela (2018). "Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry". In: *International Conference on Machine Learning*.

Nickel, M. and D. Kiela (2017). "Poincaré embeddings for learning hierarchical representations". In: *Advances in neural information processing systems*, pp. 6338–6347.

Pennington, J., R. Socher, and C. Manning (2014). "Glove: Global Vectors for Word Representation". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Peyré, G. and M. Cuturi (2018). *Computational Optimal Transport*. Tech. rep.

Pradet, Q., G. De Chalendar, and J. Baguenier-Desormeaux (2014). "WoNeF, an improved, expanded and evaluated automatic French translation of WordNet". In: *Proceedings of the Seventh Global Wordnet Conference*, pp. 32–39.

Rangarajan, A., H. Chui, and F. L. Bookstein (1997). "The Softassign Procrustes Matching Algorithm". In: *Lecture Notes in Computer Science* 1230, pp. 29–42.

Rossi, R. A. and N. K. Ahmed (2015). "The Network Data Repository with Interactive Graph Analytics and Visualization". In: *AAAI*.

Rubner, Y., C. Tomasi, and L. J. Guibas (2000). "Earth mover's distance as a metric for image retrieval". In: *International Journal of Computer Vision*.

Rusinkiewicz, S. and M. Levoy (2001). "Efficient variants of the ICP algorithm." In: *3DIM*. Vol. 1, pp. 145–152.

Sala, F., C. De Sa, A. Gu, and C. Re (2018). "Representation Tradeoffs for Hyperbolic Embeddings". In: *International Conference on Machine Learning*, pp. 4457–4466.

Salimans, T., H. Zhang, A. Radford, and D. Metaxas (2018). "Improving GANs using optimal transport". In: *International Conference on Learning Representations*.

Santambrogio, F. (2010). "Introduction to Optimal Transport Theory". In: *arXiv preprint arXiv:1009.3856*, pp. 1–16.

Saveski, M. and I. Trajkovski (2010). "Automatic construction of wordnets by using machine translation and language modeling". In: *Proceedings of the Seventh Language Technologies Conference*.

Shi, L. and R. Mihalcea (2005). "Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing". In: *International conference on intelligent text processing and computational linguistics*. Springer, pp. 100–111.

Shvaiko, P. and J. Euzenat (2013). *Ontology matching: State of the art and future challenges*.

Spohr, D., L. Hollink, and P. Cimiano (2011). "A machine learning approach to multilingual and cross-lingual ontology matching". In: *International Semantic Web Conference*. Springer, pp. 665–680.

Tay, Y., L. A. Tuan, and S. C. Hui (2018). "Hyperbolic representation learning for fast and efficient neural question answering". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, pp. 583–591.

Vayer, T., L. Chapel, R. Flamary, R. Tavenard, and N. Courty (2018). "Fused Gromov-Wasserstein distance for structured objects: theoretical foundations and mathematical properties". In: *arXiv preprint arXiv:1811.02834*.

Villani, C. (2008). *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

Wilson, B. and M. Leimeister (2018). "Gradient descent in hyperbolic space". In: *arXiv preprint arXiv:1805.08207*.

Zhang, H., S. J. Reddi, and S. Sra (2016). "Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds". In: *Advances in Neural Information Processing Systems*, pp. 4592–4600.

Zhang, M., Y. Liu, H. Luan, and M. Sun (2017). "Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1934–1945.

A Pretraining Strategies

The loss function (10) is highly non-convex with respect to θ, a consequence of both the objective itself and the nature of hyperbolic neural networks (Ganea et al., 2018b). As a result, we found that initialization plays a crucial role in this problem, since it is very hard to overcome a poor initial local minimum. Even layer-wise random initialization of weights and biases proved futile. As a solution, we experimented with the following three pre-training initialization schemes, all of which intuitively try to approximately ensure (in different ways) that f does not "collapse" the space $\mathcal{Y}$:

Identity. Initialize f_θ to approximate the identity:

$$\min_\theta \sum_{i=1}^n d_{\mathbb{D}}(\mathbf{y}_i, f_\theta(\mathbf{y}_i)),$$

which trivially ensures that f_θ (approximately) preserves the overall geometry of the space.

CrossMap. Initialize f_θ to approximately match the target points to the source points in a random permuted order:

$$\min_\theta \sum_{i=1}^n d_{\mathbb{D}}(\mathbf{x}_{\sigma(i)}, f_\theta(\mathbf{y}_i))$$

for some permutation $\sigma(i)$, which again ensures that f_θ approximately preserves the global geometry, albeit for an arbitrary labeling of the points.

Procrustes. Following (Bunne et al., 2019), we initialize f_θ to be approximately end-to-end orthogonal:

$$\min_\theta \sum_{i=1}^n d_{\mathbb{D}}(f(\mathbf{y}_i), \mathbf{P}\mathbf{y}_i),$$

where $\mathbf{P} = \operatorname{argmin}_{\mathbf{P}\in O(n)} \|\mathbf{X} - \mathbf{P}\mathbf{Y}\|_2^2$, i.e., $\mathbf{P}$ is the solution of (a hyperbolic version of) the Orthogonal Procrustes problem for mapping $\mathbf{Y}$ to $\mathbf{X}$, which can be obtained via singular value decomposition (SVD). This strategy thus requires computing an SVD for every gradient update on θ; hence, it is significantly more computationally expensive than the other two.

B Optimization Details

Each forward pass of the loss function (10) requires solving three regularized OT problems. While this can be done to completion in $O(N^2 \log N\varepsilon^{-3})$ time (Altschuler et al., 2017), practical implementations often run the Sinkhorn algorithm for a fixed number of iterations with a tolerance threshold on the objective improvement. We rely on the `geomloss`[3] package for efficient differentiable Sinkhorn divergence implementation and on the `geoopt`[4] package for Riemannian optimization. We run our method for a fixed number of outer iterations (200 in all our experiments), which given the decay strategy on the entropy regularization parameter ε, ensures that ε ranges from 1×10^1 to 1×10^{-2}. All experiments where run a single machine with 32-core processor, Intel Xeon CPU @3.20 GHz, and exploiting computations on the GPU (a single GeForce Titan X) whenever possible. With this configuration the total runtime of our method on the experiments ranged from < 1 to 20 minutes.

C Dataset Details

To generate the parallel WordNet datasets, we use the `nltk` interface to WordNet, and proceed as follows. In the English WordNet, we first filter out all words except nouns, and generate their transitive closure. For each of the remaining synsets, we query for lemmas in each of the four other languages (Es, Fr, It, Ca), for which `nltk` provides multilingual support in WordNet. These tuples of lemmas form our ground-truth translations, which are eventually split into a validation set of size 5000, leaving all the other pairs for test data (approximately 1500 for each language pairs). Note that the validation is for visualization purposes only, and all model selection is done in a purely unsupervised way based on the training objective. After the multi-lingual synset vocabularies have been extracted, we ensure their transitive closures are complete and write all the relations in these closures to a file, which will be used as an input to the PoincareEmbeddings toolkit.[5]

To generate the datasets for the synthetic noise-sensitivity experiments (§7.2), we start from the original CS-PhD dataset.[6] Given a pre-defined value ν, we iterate through the hierarchy removing node x with probability p, connecting x's children with x's parent to keep the tree connected. We repeat this with noise values $p \in \mathfrak{P} = [0.01, 0.05, 0.1, 0.2]$ and embed all of these using the PoincareEmbeddings in hyperbolic spaces of dimensions $d \in \mathfrak{D} = [2, 5, 10, 20]$. For a given dimensionality and noise level, we use our method to find correspondences between the noise-less and noisy version of the hierarchy (i.e., $|\mathfrak{P}| \times |\mathfrak{N}|$ matching tasks in total).

Statistics about all the datasets used in this work are provided in Table 3. Further details about the OAEI datasets can found on the project's website.[7]

[3] https://www.kernel-operations.io/geomloss/

[4] https://geoopt.readthedocs.io/en/latest/

[5] https://github.com/facebookresearch/poincare-embeddings

[6] http://networkrepository.com/CSphd.php

[7] http://oaei.ontologymatching.org/2018/

	WordNet				Anatomy		Biodiv			
	English (En)	Spanish (Es)	French (Fr)	Catalan (Ca)	Human	Mouse	Flopo	Pto	Envo	Sweet
Entities	8206	8206	8206	8206	3298	2737	360	1456	6461	4365
Relations	47938	47938	47938	47938	18556	7364	472	11283	73881	30101
Embedding Size	10	10	10	10	10	10	10	10	10	10

Table 3: Dataset characteristics. All datasets embedded with the method of Nickel and Kiela (2017).

Model	Metric	Cost	Pretrain	ε-annealing	Layers	Hidden dim	Layer Type	Nonlin.	Opt	LR
Full	Poincare	$d(x,y)$	CrossMap	$10^1 \to 10^{-2}$	10	20	HyperLinear	elu	Radam	10^{-3}
Small	–	–	–	–	2	10	–	–	–	–
Euclidean	Euclidean	–	–	–	–	–	–	–	–	–
ReLU	–	–	–	–	–	–	ReLU	–	–	
Rsgd	–	–	–	–	–	–	–	–	Rsgd	5×10^{-2}
Möbius	–	–	–	–	–	10	Möbius	–	–	–
cosh Cost	–	$-\cosh \circ d_{\mathbb{D}}$	–	–	–	–	–	–	–	–
No Pretrain	–	–	None	–	–	–	–	–	–	–

Table 4: Ablated model configurations for the monolingual En →En WordNet task.

D Model Configurations and Hyperparameters

In Table 4, we provide full configuration details for all the ablated models used in the WordNet En →En self-recovery experiment (results shown in Table 1a). Dashed lines indicate a parameter being the same as in the Full Model.

E A Brief Summary of Theoretical Guarantees for Optimal Transport (Euclidean Case)

As mentioned in Section 4.2, whenever optimal transport is used with the goal of obtaining correspondences, there are various theoretical considerations that become particularly appealing.

The first of such considerations pertains to the nature of the solution, i.e., the optimal coupling π^* which minimizes the cost (5). When the final end goal is to transport points from one space to the other, the best case scenario would be if the optimal π happens to be a "hard" deterministic mapping. A celebrated result by Brenier (1987) (see also (Brenier, 1991)) shows that this indeed the case for the quadratic cost,[8] i.e., for the 2-Wasserstein distance. Even when solving the problem approximately with entropic regularization (cf. Eq. (7)), this result guarantees that the solution found in this way converges to a deterministic mapping as $\varepsilon \to 0$.

Now, assuming now that such a map exists, the next aspect we might be interested in is its smoothness. Intuitively, smoothness of this mapping is desirable since it is more likely to lead to robust matchings in the context of correspondences, even if, again, the argument holds asymptotically for the regularized problem. This, clearly, is a very strong property to require. While not even continuity can be guaranteed in general (Ambrosio and Gigli, 2013), again for the quadratic-cost things are simpler: if the source and target densities are smooth and the support of the target distribution satisfies suitable convexity assumptions, the optimal map is guaranteed to be smooth too (Caffarelli, 1992a; Caffarelli, 1992b).

[8] This result holds in more general settings. We refer the reader to (Santambrogio, 2010; Ambrosio and Gigli, 2013) for further details.

F A Brief Summary of Theoretical Guarantees for Optimal Transport (Riemannian Manifold Case)

Extending the problem beyond Euclidean to more general spaces has been one of the central questions theoretical optimal transport research over the past decades (Villani, 2008). For obvious reasons, here we focus the discussion on results related to hyperbolic spaces, and more generally, to Riemannian manifolds.

Let us first note that Problem (5) is well-defined for any complete and separable metric space $\mathcal{X}$. Since the arclength metric of a Riemannian manifold allows for the direct construction of an accompanying metric space $(\mathcal{X}, d_{\mathcal{X}})$, then OT can be defined over those too. However, some of the theoretical results of their Euclidean counterparts do not transfer that easily to the Riemannian case (Ambrosio and Gigli, 2013). Nevertheless, the existence and uniqueness of the optimal transportation plan π^*, which in addition is induced by

a transport map T, can be guaranteed with mild regularity conditions on the source distribution α. This was first shown in seminal work by McCann (2001). The result, which acts as an Riemannian analogue of that of Brenier for the Euclidean setting (Brenier, 1987), is shown below as presented by Ambrosio and Gigli (2013):

Theorem F.1 (McCann, version of (Ambrosio and Gigli, 2013))**.** *Let M be a smooth, compact Riemannian manifold without boundary and $\alpha \in \mathcal{P}(\mathcal{M})$. Then the following are equivalent:*

(i) $\forall \beta \in \mathcal{P}(\mathcal{M})$, there exists a unique optimal $\pi \in \Pi(\alpha, \beta)$, and this plan is induced by a map T.

(ii) α is regular.

If either (i) or (ii) holds, the optimal T can be written as $x \mapsto \exp_x(-\nabla\phi(x))$ for some c-concave function $\phi : \mathcal{M} \to \mathbb{R}$.

The question of regularity of the optimal map, on the other hand, is much more delicate now than in the Euclidean case (Ambrosio and Gigli, 2013; Ma et al., 2005; Loeper, 2009). In addition to the suitable convexity assumptions on the support of the target density, a restrictive structural condition, known as the Ma-Trudinger-Wang (MTW) condition (Ma et al., 2005), needs to be imposed on the cost in order to guarantee continuity of the optimal map. Unfortunately for our setting, in the case of Riemannian manifolds the MTW condition for the usual quadratic cost $c = d^2/2$ is so restrictive that it implies that $\mathcal{X}$ has non-negative sectional curvature (Loeper, 2009), which rules out hyperbolic spaces. However, a recent sequence of remarkable results (Lee and Li, 2012; Li, 2009) prove that for simple variations of the Riemannian metric d on hyperbolic spaces, smoothness is again guaranteed:

Theorem F.2 (Lee and Li, (Lee and Li, 2012))**.** *Let d be the Riemannian distance function on a manifold of constant sectional curvature -1; then the cost functions $-\cosh \circ d$ and $-\log \circ (1 + \cosh) \circ d$ satisfy the strong MTW condition, and the cost functions $\pm \log \circ \cosh \circ d$ satisfy the weak MTW condition.*

Thus, these cost objectives can be used it out hyperbolic optimal transport matching setting with the hopes of obtaining a smoother solution, and therefore a more stable set of correspondences.