Learning Multilingual Word Embeddings in Latent Metric Space: A Geometric Approach

Pratik Jawanpuria[1], **Arjun Balgovind**[2*], **Anoop Kunchukuttan**[1], **Bamdev Mishra**[1]

[1]Microsoft, India [2]IIT Madras, India
[1]{pratik.jawanpuria,ankunchu,bamdevm}@microsoft.com
[2]barjun@cse.iitm.ac.in

Abstract

We propose a novel geometric approach for learning bilingual mappings given monolingual embeddings and a bilingual dictionary. Our approach decouples the source-to-target language transformation into (a) language-specific rotations on the original embeddings to align them in a common, latent space, and (b) a language-independent similarity metric in this common space to better model the similarity between the embeddings. Overall, we pose the bilingual mapping problem as a classification problem on smooth Riemannian manifolds. Empirically, our approach outperforms previous approaches on the bilingual lexicon induction and cross-lingual word similarity tasks.

We next generalize our framework to represent multiple languages in a common latent space. Language-specific rotations for all the languages and a common similarity metric in the latent space are learned *jointly* from bilingual dictionaries for multiple language pairs. We illustrate the effectiveness of joint learning for multiple languages in an indirect word translation setting.

1 Introduction

Bilingual word embeddings are a useful tool in natural language processing (NLP) that has attracted a lot of interest lately due to a fundamental property: similar concepts/words across different languages are mapped close to each other in a common embedding space. Hence, they are useful for joint/transfer learning and sharing annotated data across languages in different NLP applications such as machine translation (Gu et al., 2018), building bilingual dictionaries (Mikolov et al., 2013b), mining parallel corpora (Conneau et al., 2018), text classification (Klementiev et al., 2012), sentiment analysis (Zhou et al., 2015), and dependency parsing (Ammar et al., 2016).

Mikolov et al. (2013b) empirically show that a linear transformation of embeddings from one language to another preserves the geometric arrangement of word embeddings. In a supervised setting, the transformation matrix, $\mathbf{W}$, is learned given a small bilingual dictionary and their corresponding monolingual embeddings. Subsequently, many refinements to the bilingual mapping framework have been proposed (Xing et al., 2015; Smith et al., 2017b; Conneau et al., 2018; Artetxe et al., 2016, 2017, 2018a,b).

In this work, we propose a novel geometric approach for learning bilingual embeddings. We rotate the source and target language embeddings from their original vector spaces to a common latent space via language-specific orthogonal transformations. Furthermore, we define a similarity metric, the Mahalanobis metric, in this common space to refine the notion of similarity between a pair of embeddings. We achieve the above by learning the transformation matrix as follows: $\mathbf{W} = \mathbf{U}_t \mathbf{B} \mathbf{U}_s^\top$, where $\mathbf{U}_t$ and $\mathbf{U}_s$ are the orthogonal transformations for target and source language embeddings, respectively, and $\mathbf{B}$ is a positive definite matrix representing the Mahalanobis metric.

*This work was carried out during the author's internship at Microsoft, India.

Transactions of the Association for Computational Linguistics, vol. 7, pp. 107–120, 2019. Action Editor: Doug Downey.
Submission batch: 10/2018; Revision batch: 11/2018; Published 4/2019.
© 2019 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

The proposed formulation has the following benefits:

- The learned similarity metric allows for a more effective similarity comparison of embeddings based on evidence from the data.
- A common latent space decouples the source and target language transformations, and naturally enables representation of word embeddings from both languages in a single vector space.
- We also show that the proposed method can be easily generalized to jointly learn multilingual embeddings, given bilingual dictionaries of multiple language pairs. We map multiple languages into a single vector space by learning the characteristics common across languages (the similarity metric) as well as language-specific attributes (the orthogonal transformations).

The optimization problem resulting from our formulation involves orthogonal constraints on language-specific transformations ($\mathbf{U}_i$ for language i) as well as the symmetric positive-definite constraint on the metric $\mathbf{B}$. Instead of solving the optimization problem in the Euclidean space with constraints, we view it as an optimization problem in smooth Riemannian manifolds, which are well-studied topological spaces (Lee, 2003). The Riemannian optimization framework embeds the given constraints into the search space and conceptually views the problem as an unconstrained optimization problem over the manifold.

We evaluate our approach on different bilingual as well as multilingual tasks across multiple languages and datasets. The following is a summary of our findings:

- Our approach outperforms state-of-the-art supervised and unsupervised bilingual mapping methods on the bilingual lexicon induction as well as the cross-lingual word similarity tasks.
- An ablation analysis reveals that the following contribute to our model's improved performance: (a) aligning the embedding spaces of different languages, (b) learning a similarity metric which induces a latent space, (c) performing inference in the induced latent space, and (d) formulating the tasks as a classification problem.
- We evaluate our multilingual model on an indirect word translation task: translation between a language pair that does not have a bilingual dictionary, but the source and target languages each possess a bilingual dictionary with a third, common pivot language. Our multilingual model outperforms a strong unsupervised baseline as well as methods based on adapting bilingual methods for this indirect translation task.
- Lastly, we propose a semi-supervised extension of our approach that further improves performance over the supervised approaches.

The rest of the paper is organized as follows. Section 2 discusses related work. The proposed framework, including problem formulations for bilingual and multilingual mappings, is presented in Section 3. The proposed Riemannian optimization algorithm is described in Section 4. In Section 5, we discuss our experimental setup. Section 6 presents the results of experiments on direct translation with our algorithms and analyzes the results. Section 7 presents experiments on indirect translation using our generalized multilingual algorithm. We discuss a semi-supervised extension to our framework in Section 8. Section 9 concludes the paper.

2 Related Work

Bilingual Embeddings. Mikolov et al. (2013b) show that a linear transformation from embeddings of one language to another can be learned from a bilingual dictionary and corresponding monolingual embeddings by performing linear least-squares regression. A popular modification to this formulation constrains the transformation matrix to be orthogonal (Xing et al., 2015; Smith et al., 2017b; 2018a). This is known as the *orthogonal Procrustes problem* (Schönemann, 1966). Orthogonality preserves monolingual distances and ensures the transformation is reversible. Lazaridou et al. (2015) and Joulin et al. (2018) optimize alternative loss functions in this framework. Artetxe et al. (2018a) improves on the Procrustes solution and propose a multi-step framework consisting of a series of linear transformations to the data. Faruqui and Dyer (2014)

use Canonical Correlation Analysis (CCA) to learn linear projections from the source and target languages to a common space such that correlations between the embeddings projected to this space are maximized. Procrustes solution–based approaches have been shown to perform better than CCA-based approaches (Artetxe et al., 2016, 2018a).

We view the problem of mapping the source and target languages word embeddings as (a) aligning the two language spaces and (b) learning a similarity metric in this (learned) common space. We accomplish this by learning suitable language-specific orthogonal transformations (for alignment) and a symmetric positive-definite matrix (as Mahalanobis metric). The similarity metric is useful in addressing the limitations of mapping to a common latent space under orthogonality constraints, an issue discussed by Doval et al. (2018). Whereas Doval et al. (2018) learn a second correction transformation by assuming the average of the projected source and target embeddings as the true latent representation, we make no such assumption and learn the similarity metric from the data. Kementchedjhieva et al. (2018), recently, employed the generalized Procrustes analysis (GPA) method (Gower, 1975) for the bilingual mapping problem. GPA maps both the source and target language embeddings to a latent space, which is constructed by averaging over the two language spaces.

Unsupervised methods have shown promising results, matching supervised methods in many studies. Artetxe et al. (2017) proposed a bootstrapping method for bilingual lexicon induction problem by using a small-seed bilingual dictionary. Subsequently, Artetxe et al. (2018b) and Hoshen and Wolf (2018) have proposed initialization methods that eliminate the need for a seed dictionary. Zhang et al. (2017b) and Grave et al. (2018) proposed aligning the source and target language word embeddings by optimizing the Wasserstein distance. Unsupervised methods based on adversarial training objectives have also been proposed (Barone, 2016; Zhang et al., 2017a; Conneau et al., 2018; Chen and Cardie, 2018). A recent work by Søgaard et al. (2018) discusses cases in which unsupervised bilingual lexicon induction does not lead to good performance.

Multilingual Embeddings. Ammar et al. (2016) and Smith et al. (2017a) adapt bilingual approaches for representing embeddings of multiple languages in a common vector space by designating one of the languages as a *pivot* language. In this simple approach, bilingual mappings are learned *independently* from all other languages to the pivot language. A GPA-based method (Kementchedjhieva et al., 2018) may also be used to jointly transform multiple languages to a common latent space. However, this requires an n-way dictionary to represent n languages. In contrast, the proposed approach requires only pairwise bilingual dictionaries such that every language under consideration is represented in at least one bilingual dictionary.

The above-mentioned approaches are referred to as *offline* since the monolingual and bilingual embeddings are learned separately. In contrast, *online* approaches directly learn a bilingual/ multilingual embedding from parallel corpora (Hermann and Blunsom, 2014; Huang et al., 2015; Duong et al., 2017), optionally augmented with monolingual corpora (Klementiev et al., 2012; Chandar et al., 2014; Gouws et al., 2015). In this work, we focus on offline approaches.

3 Learning Latent Space Representation

In this section, we first describe the proposed geometric framework to learn bilingual embeddings. We then present its generalization to the multilingual setting.

3.1 Geometry-aware Factorization

We propose to transform the word embeddings from the source and target languages to a common space in which the similarity of word embeddings may be better learned. To this end, we *align* the source and target languages embedding spaces by learning language-specific rotations: $\mathbf{U}_s \in \mathbb{O}^d$ and $\mathbf{U}_t \in \mathbb{O}^d$ for the source and target languages embeddings, respectively. Here $\mathbb{O}^d$ represents the space of d-dimensional orthogonal matrices. An embedding x in the source language is thus transformed to $\psi_s(x) = \mathbf{U}_s^\top x$. Similarly, for an embedding z in the target language: $\psi_t(z) = \mathbf{U}_t^\top z$. These orthogonal transformations map (align) both the source and target language embeddings to a common space in which we learn a data-dependent similarity measure, as discussed below.

We learn a Mahalanobis metric $\mathbf{B}$ to *refine* the notion of similarity[1] between the two transformed embeddings $\psi_s(x)$ and $\psi_t(z)$. The Mahalanobis metric incorporates the feature correlation information from the given training data. This allows for a more effective similarity comparison of language embeddings (than the cosine similarity). In fact, Mahalanobis similarity measure reduces to cosine similarity when the features are uncorrelated and have unit variance, which may be a strong assumption in real-world applications. Søgaard et al. (2018) have argued that monolingual embedding spaces across languages are not necessarily isomorphic, hence learning an orthogonal transformation alone may not be sufficient. A similarity metric learned from the data may mitigate this limitation to some extent by learning a correction in the latent space.

Since $\mathbf{B}$ is a Mahalanobis metric in $\mathbb{R}^d$ space, it is a $d \times d$ symmetric positive-definite matrix $\mathbf{B}$, *i.e.,* $\mathbf{B} \succ \mathbf{0}$. The similarity between the embeddings x and z in the proposed setting is computed as $h_{st}(x, z) = \psi_t(z)^\top \mathbf{B} \psi_s(x) = z^\top (\mathbf{U}_t \mathbf{B} \mathbf{U}_s^\top) x$. The source to the target language transformation is expressed as $\mathbf{W}_{ts} = \mathbf{U}_t \mathbf{B} \mathbf{U}_s^\top$. For an embedding x in the source language, its transformation to the target language space is given by $\mathbf{W}_{ts} x$.

The proposed factorization of the transformation $\mathbf{W} = \mathbf{U}\mathbf{B}\mathbf{V}^\top$, where $\mathbf{U}, \mathbf{V} \in \mathbb{O}^d$ and $\mathbf{B} \succ \mathbf{0}$, is sometimes referred to as polar factorization of a matrix (Bonnabel and Sepulchre, 2010; Meyer et al., 2011). Polar factorization is similar to the singular value decomposition (SVD) The key difference is that SVD enforces $\mathbf{B}$ to be a *diagonal* matrix with non-negative entries, which accounts for only the axis rescaling instead of full feature correlation and is more difficult to optimize (Mishra et al., 2014; Harandi et al., 2017).

3.2 Latent Space Interpretation

Computing the Mahalanobis similarity measure is equivalent to computing the cosine similarity in a special latent (feature) space. This latent space is defined by the transformation $\phi : \mathbb{R}^d \to \mathbb{R}^d$, where the mapping is defined as $\phi(w) = \mathbf{B}^{\frac{1}{2}} w$. Since $\mathbf{B}$ is a symmetric positive-definite matrix, $\mathbf{B}^{\frac{1}{2}}$ is well-defined and unique.

Hence, our model may equivalently be viewed as learning a suitable latent space as follows. The source and target languages embeddings are linearly transformed as $x \mapsto \phi(\psi_s(x))$ and $z \mapsto \phi(\psi_t(z))$, respectively. The functions $\phi(\psi_s(\cdot))$ and $\phi(\psi_t(\cdot))$ map the source and target language embeddings, respectively, to a common latent space. We learn the matrices $\mathbf{B}$, $\mathbf{U}_s$, and $\mathbf{U}_t$ corresponding to the transformations $\phi(\cdot)$, $\psi_s(\cdot)$, and $\psi_t(\cdot)$, respectively. Since the matrix $\mathbf{B}$ is embedded implicitly in this latent feature space, we employ the usual cosine similarity measure, computed as $\phi(\psi_t(z))^\top \phi(\psi_s(x)) = z^\top \mathbf{U}_t \mathbf{B} \mathbf{U}_s^\top x$. It should be noted that this is equal to $h_{st}(x, z)$.

3.3 A Classification Model

We assume a small bilingual dictionary (of size n) is available as the training data. Let $\mathbf{X}_s \in \mathbb{R}^{d \times n_s}$ and $\mathbf{X}_t \in \mathbb{R}^{d \times n_t}$ denote the embeddings of the dictionary words from the source and target languages, respectively. Here, n_s and n_t are the number of unique words in the source and target languages present in the dictionary.

We propose to model the bilingual word embedding mapping problem as a binary classification problem. Consider word embeddings x and z from the source and target languages, respectively. If the words corresponding to x and z constitute a translation pair then the pair $\{x, z\}$ belongs to the positive class, else it belongs to the negative class. The prediction function for the pair $\{x, z\}$ is $h_{st}(x, z)$. We create a binary label matrix $\mathbf{Y}_{st} \in \{0, 1\}^{n_s \times n_t}$ whose (i, j)-th entry corresponds to the correctness of mapping the i-th embedding in $\mathbf{X}_s$ to the j-th embedding in $\mathbf{X}_t$. Our overall optimization problem is as follows:

$$\min_{\mathbf{U}_s \in \mathbb{O}^d, \mathbf{U}_t \in \mathbb{O}^d, \mathbf{B} \succ \mathbf{0}} \|\mathbf{X}_s^\top \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top \mathbf{X}_t - \mathbf{Y}_{st}\|_F^2 + \lambda \|\mathbf{B}\|_F^2. \quad (1)$$

where $\|\cdot\|_F$ is the Frobenius norm and $\lambda > 0$ is the regularization parameter. We employ the square loss function since it is smooth and relatively easier to optimize. It should be noted that our prediction function is invariant of the direction of mapping, *i.e.,* $h_{st}(x, z) = h_{ts}(z, x)$. Hence, our model learns bidirectional mapping. The transformation matrix from the target to the source language is given by $\mathbf{W}_{st} = \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top$, *i.e.,* $\mathbf{W}_{st} = \mathbf{W}_{ts}^\top$.

[1] Mahalanobis metric generalizes the notion of cosine similarity. For given two unit normalized vectors $x_1, x_2 \in \mathbb{R}^d$, their cosine similarity is given by $\text{sim}_{\mathbf{I}}(x_1, x_2) = x_1^\top \mathbf{I} x_2 = x_1^\top x_2$, where $\mathbf{I}$ is the identity matrix. If this space is endowed with a metric $\mathbf{B} \succ \mathbf{0}$, then $\text{sim}_{\mathbf{B}}(x_1, x_2) = x_1^\top \mathbf{B} x_2$.

The computation complexity of computing the loss term in (1) is linear in n, the size of the given bilingual dictionary. This is because the loss term in (1) can be re-written as follows:

$$\begin{aligned}&\|\mathbf{X}_s^\top \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top \mathbf{X}_t - \mathbf{Y}_{st}\|_F^2 \\ &= \mathrm{Tr}\big(\mathbf{U}_t \mathbf{B} \mathbf{U}_s^\top (\mathbf{X}_s \mathbf{X}_s^\top) \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top (\mathbf{X}_t \mathbf{X}_t^\top)\big) + |\Omega| \\ &\quad - 2 \sum_{\{(i,j):(i,j)\in\Omega\}} x_{si}^\top \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top x_{tj}, \end{aligned} \tag{2}$$

where x_{si} represents the i-th column in $\mathbf{X}_s$, x_{tj} represents the j-th column in $\mathbf{X}_t$, Ω is the set of row-column indices corresponding to entry value 1 in $\mathbf{Y}_{st}$, and $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. The complexity of computing the first and third term in (2) is $O(d^3 + n_s d^2 + n_t d^2)$ and $O(nd + n_s d^2 + n_t d^2)$, respectively. Similarly, the computation cost of the gradient of the objective function in (1) is also linear in n. Hence, our framework can efficiently leverage information from all the negative samples.

In the next section, we discuss a generalization of our approach to multilingual settings.

3.4 Generalization to Multilingual Setting

In this section, we propose a unified framework for learning mappings when bilingual dictionaries are available for multiple language pairs. We formalize the setting as an undirected, connected graph $G(V, E)$, where each node represents a language and an edge represents the availability of a bilingual dictionary between the corresponding pair of languages. Given all bilingual dictionaries corresponding to the edge set E, we propose to align the embedding spaces of all languages in the node set V and learn a common latent space for them.

To this end, we *jointly* learn an orthogonal transformation $\mathbf{U}_i \in \mathbb{O}^d$ for every language L_i and the Mahalanobis metric $\mathbf{B} \succ \mathbf{0}$. The latter is *common* across all languages in the multilingual setup and helps incorporate information across languages in the latent space. It should be noted that the transformation $\mathbf{U}_i$ is employed for all the bilingual mapping problems in this graph associated with L_i. The transformation from L_i to L_j is given by $\mathbf{W}_{ji} = \mathbf{U}_j \mathbf{B} \mathbf{U}_i^\top$. Further, we are also able to obtain transformations between any language pair in the graph, even if a bilingual dictionary between them is not available.

Let $\mathbf{X}_i^j \in \mathbb{R}^{d\times m}$ be[2] the embeddings of the dictionary words of L_i in the dictionary corresponding to edge $e_{ij} \in E$. Let $\mathbf{Y}_{ij} \in \{0,1\}^{m\times m}$ be the binary label matrix corresponding to the dictionary between L_i and L_j. The proposed optimization problem for multilingual setting is

$$\begin{aligned}\min_{\substack{\mathbf{U}_i\in\mathbb{O}^d\ \forall i \\ \mathbf{B}\succ\mathbf{0}}} \sum_{e_{ij}\in E} \frac{1}{|\Omega_{ij}|} \|(\mathbf{X}_i^j)^\top \mathbf{U}_i \mathbf{B} \mathbf{U}_j^\top \mathbf{X}_j^i - \mathbf{Y}_{ij}\|_F^2 \\ + \lambda \|\mathbf{B}\|_F^2. \end{aligned} \tag{3}$$

We term our approach as **Geo**metry-aware **M**ultilingual **M**apping (GeoMM). We next discuss the optimization algorithm for solving the bilingual mapping problem (1) as well as its generalization to the multilingual setting (3).

4 Optimization Algorithm

The geometric constraints $\mathbf{U}_s \in \mathbb{O}^d, \mathbf{U}_t \in \mathbb{O}^d$, and $\mathbf{B} \succ \mathbf{0}$ in the proposed problems (1) and (3) have been studied as smooth Riemannian manifolds, which are well explored topological spaces (Edelman et al., 1998). The orthogonal matrices $\mathbf{U}_i$ lie in, what is popularly known as, the d-dimensional Orthogonal manifold. The space of $d \times d$ symmetric positive definite matrices ($\mathbf{B} \succ \mathbf{0}$) is known as the Symmetric Positive Definite manifold. The Riemannian optimization framework embeds such constraints into the search space and conceptually views the problem as an unconstrained problem over the manifolds. In the process, it is able to exploit the geometry of the manifolds and the symmetries involved in them. Absil et al. (2008) discuss several tools to systematically optimize such problems. We optimize the problems (1) and (3) using the Riemannian conjugate gradient algorithm (Absil et al., 2008; Sato and Iwai, 2013).

Publicly available toolboxes such as Manopt (Boumal et al., 2014), Pymanopt (Townsend et al., 2016), or ROPTLIB (Huang et al., 2016) have scalable off-the-shelf generic implementations of several Riemannian optimization algorithms. We employ Pymanopt in our experiments, where we only need to supply the objective function.

[2]For notational convenience, the number of unique words in every language in all their dictionaries is kept same (m).

5 Experimental Settings

In this section, we describe the evaluation tasks, the datasets used, and the experimental details of the proposed approach.

Evaluation Tasks. We evaluate our approach on several tasks:

- To evaluate the quality of the bilingual mappings generated, we evaluate our algorithms primarily for the bilingual lexicon induction (BLI) task, *i.e.,* word translation task, and compare Precision@1 with previously reported state-of-the-art results on benchmark datasets (Dinu and Baroni, 2015; Artetxe et al., 2016; Conneau et al., 2018).
- We also evaluate on the cross-lingual word similarity task using the SemEval 2017 dataset.
- To ensure that quality of embeddings on monolingual tasks does not degrade, we evaluate the quality of our embeddings on the monolingual word analogy task (Artetxe et al., 2016).
- To illustrate the utility of representing embeddings of multiple language in a single latent space, we evaluate our multilingual embeddings on the one-hop translation task, *i.e.,* a direct dictionary between the source and target languages is not available, but the source and target languages share a bilingual dictionary with a pivot language.

Datasets. For bilingual and multilingual experiments, we report results on the following widely used, publicly available datasets:

- **VecMap**: This dataset was originally made available by Dinu and Baroni (2015) with subsequent extensions by other researchers (Artetxe et al., 2017, 2018a). It contains bilingual dictionaries from English (en) to four languages: Italian (it), German (de), Finnish (fi), and Spanish (es). The detailed experimental settings for this BLI task can be found in Artetxe et al. (2018b).
- **MUSE**: This dataset was originally made available by Conneau et al. (2018). It contains bilingual dictionaries from English to many languages such as Spanish (es), French (fr), German (de), Russian (ru), Chinese (zh), and *vice versa*. The detailed experimental settings for this BLI task can be found in Conneau et al. (2018). This dataset also contains bilingual dictionaries between several other European languages, which we employ in multilingual experiments.

Experimental Settings of GeoMM. We select the regularization hyper-parameter λ from the set $\{10, 10^2, 10^3, 10^4\}$ by evaluation on a validation set created out of the training dataset. For inference, we use the (normalized) latent space representations of embeddings ($\mathbf{B}^{\frac{1}{2}}\mathbf{U}_i^\top x$) to compute similarity between the embeddings. For inference in the bilingual lexicon induction task, we employ the Cross-domain Similarity Local Scaling (CSLS) similarity score (Conneau et al., 2018) in nearest neighbor search, unless otherwise mentioned. CSLS has been shown to perform better than other methods in mitigating the *hubness* problem (Dinu and Baroni, 2015) for search in high-dimensional spaces.

While discussing experiments, we denote our bilingual mapping algorithm (Section 3.3) as GeoMM and its generalization to the multilingual setting (Section 3.4) as GeoMM$_{\text{multi}}$. Our code is available at `https://github.com/anoopkunchukuttan/geomm`.

6 Direct Translation: Results and Analysis

In this section, we evaluate the performance of our approach on two tasks: bilingual lexicon induction and cross-lingual word similarity. We also perform ablation tests to understand the effect of major sub-components of our algorithm. We verify the monolingual performance of the mapped embeddings generated by our algorithm.

6.1 Bilingual Lexicon Induction (BLI)

We compare GeoMM with the best performing supervised methods. We also compare with unsupervised methods as they have been shown to be competitive with supervised methods. The following baselines are compared in the BLI experiments.

- Procrustes: the bilingual mapping is learned by solving the orthogonal Procrustes problem (Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017b; Conneau et al., 2018).
- MSF: the Multi-Step Framework proposed by Artetxe et al. (2018a), with CSLS retrieval.

Method	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	avg.
Supervised											
GeoMM	81.9	85.5	82.1	**84.2**	74.9	**76.7**	**52.8**	**67.6**	49.1	**45.3**	**70.0**
GeoMM$_{\text{multi}}$	81.0	**85.7**	81.9	83.9	75.1	75.7	51.7	67.2	**49.4**	44.9	69.7
Procrustes	81.4	82.9	81.1	82.4	73.5	72.4	51.7	63.7	42.7	36.7	66.9
MSF-ISF	79.9	82.1	80.4	81.4	73.0	72.0	50.0	65.3	28.0	40.7	65.3
MSF	80.5	83.8	80.5	83.1	73.5	73.5	50.5	67.3	32.3	43.4	66.9
MSF$_{\mu}$	80.3	84.0	80.7	83.9	73.1	74.7	×	×	×	×	–
Unsupervised											
SL-unsup	82.3	84.7	82.3	83.6	75.1	74.3	49.2	65.6	0.0	0.0	59.7
Adv-Refine*	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	32.5	31.4	64.3
Grave et al. (2018)*	**82.8**	84.1	**82.6**	82.9	**75.4**	73.3	43.7	59.1	–	–	–
Hoshen and Wolf (2018)*	82.1	84.1	82.3	82.9	74.7	73.0	47.5	61.8	f.c.	f.c.	–
Chen and Cardie (2018)*	82.5	83.7	82.4	81.8	74.8	72.9	–	–	–	–	–

Table 1: Precision@1 for BLI on the MUSE dataset. Some notations: (a) '–' implies the original paper does not report result for the corresponding language pair, (b) 'f.c.' implies the original paper reports their algorithm failed to converge, (c) '×' implies that we could not run the authors' code successfully for the language pairs, and (d) '*' implies the results of the algorithm are reported in the original paper. The remaining results were obtained with the official implementation from the authors.

It improves on the original system (MSF-ISF) by Artetxe et al. (2018a), which employs inverted softmax function (ISF) score for retrieval.

- Adv-Refine: unsupervised adversarial training approach, with bilingual dictionary refinement (Conneau et al., 2018).

- SL-unsup: state-of-the-art self-learning (SL) unsupervised method (Artetxe et al., 2018b), employing structural similarity of the embeddings.

We also include results of the correction algorithm proposed by Doval et al. (2018) on the MSF results (referred to as MSF$_{\mu}$). In addition, we also include results of several recent works (Kementchedjhieva et al., 2018; Grave et al., 2018; Chen and Cardie, 2018; Hoshen and Wolf, 2018) on MUSE and VecMap datasets, which are reported in the original papers.

Results on MUSE Dataset: Table 1 reports the results on the MUSE dataset. We observe that our algorithm GeoMM outperforms all the supervised baselines. GeoMM also obtains significant improvements over unsupervised approaches.

The performance of the multilingual extension, GeoMM$_{\text{multi}}$, is almost equivalent to the bilingual GeoMM. This means that in spite of multiple embeddings being jointly learned and represented in a common space, its performance is still better than existing bilingual approaches. Thus, our multilingual framework is quite robust since languages from diverse language families have been embedded in the same space. This can allow downstream applications to support multiple languages without performance degradation. Even if bilingual embeddings are represented in a single vector space using a pivot language, the embedding quality is inferior compared with GeoMM$_{\text{multi}}$. We discuss more multilingual experiments in Section 7.

Method	en-it	en-de	en-fi	en-es	avg.
Supervised					
GeoMM	48.3	**49.3**	**36.1**	**39.3**	**43.3**
GeoMM$_{\text{multi}}$	**48.7**	49.1	36.0	39.0	43.2
Procrustes	44.9	46.5	33.5	35.1	40.0
MSF-ISF	45.3	44.1	32.9	36.6	39.7
MSF	47.7	47.5	35.4	38.7	42.3
MSF$_{\mu}$	48.4	47.7	34.7	38.9	42.4
GPA	45.3	48.5	31.4	–	–
CCA-NN	38.4	37.1	27.6	26.8	32.5
Unsupervised					
SL-unsup	48.1	48.2	32.6	37.3	41.6
Adv-Refine	45.2	46.8	0.4	35.4	31.9

Table 2: Precision@1 for BLI on the VecMap dataset. The results of MSF-ISF, SL-unsup, CCA-NN (Faruqui and Dyer, 2014), and Adv-Refine are reported by Artetxe et al. (2018b). CCA-NN employs nearest neighbor retrieval procedure. The results of GPA are reported by Kementchedjhieva et al. (2018).

Method	en-it	en-de	en-fi	en-es
GeoMM	**48.3**	**49.3**	**36.1**	**39.3**
(1) $\mathbf{W} \in \mathbb{R}^{d \times d}$	45.4	47.9	35.4	37.5
(2) $\mathbf{W} = \mathbf{B}$	26.3	26.3	19.5	21.2
(3) $\mathbf{W} = \mathbf{U}_t\mathbf{U}_s^\top$	13.2	16.0	8.8	11.8
(4) Targt space inf.	45.5	47.8	35.0	37.9
(5) Regression	46.8	43.3	33.9	35.4

Table 3: Ablation test results: Precision@1 for BLI on the VecMap dataset.

Results on VecMap Dataset: Table 2 reports the results on the VecMap dataset. We observe that GeoMM obtains the best performance in each language pair, surpassing state-of-the-art results reported on this dataset. GeoMM also outperforms GPA (Kementchedjhieva et al., 2018), which also learns bilingual embeddings in a latent space.

6.2 Ablation Tests

We next study the impact of different components of our framework by varying one component at a time. The results of these tests on VecMap dataset are shown in Table 3 and are discussed below.

(1) Classification with unconstrained $\mathbf{W}$. We learn the transformation $\mathbf{W}$ directly as follows:

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times d}} \lambda \|\mathbf{W}\|_F^2 + \|\mathbf{X}_s^\top \mathbf{W}^\top \mathbf{X}_t - \mathbf{Y}_{st}\|_F^2. \tag{4}$$

The performance drops in this setting compared with GeoMM, underlining the importance of the proposed factorization and the latent space representation. In addition, the proposed factorization helps GeoMM generalize to the multilingual setting ($\text{GeoMM}_{\text{multi}}$). Further, we also observe that the overall performance of this simple classification based model is better than recent supervised approaches such as Procrustes, MSF-ISF (Artetxe et al., 2018a), and GPA (Kementchedjhieva et al., 2018). This suggests that a classification model is better suited for the BLI task.

Next, we look at both components of the factorization.

(2) Without language-specific rotations. We enforce $\mathbf{U}_s = \mathbf{U}_t = \mathbf{I}$ in (1) for GeoMM, *i.e.,* $\mathbf{W} = \mathbf{B}$. We observe a significant drop in performance, which highlights the need for aligning the feature space of different languages.

(3) Without similarity metric. We enforce $\mathbf{B} = \mathbf{I}$ in (1) for GeoMM, *i.e.,* $\mathbf{W} = \mathbf{U}_t\mathbf{U}_s^\top$. It can be observed that the results are poor, which underlines the importance of a suitable similarity metric in the proposed classification model.

(4) Target space inference. We learn $\mathbf{W} = \mathbf{U}_t\mathbf{B}\mathbf{U}_s^\top$ by solving (1), as in GeoMM. During the retrieval stage, the similarity between embeddings is computed in the target space, *i.e.,* given embeddings x and z from the source and target languages, respectively, we compute the similarity of the (normalized) vectors $\mathbf{W}x$ and z. It should be noted that GeoMM computes similarity of x and z in the latent space, *i.e.,* it computes the similarity of the (normalized) vectors $\mathbf{B}^{\frac{1}{2}}\mathbf{U}_s^\top x$ and $\mathbf{B}^{\frac{1}{2}}\mathbf{U}_t^\top z$, respectively. We observe that inference in the target space degrades the performance. This shows that the latent space representation captures useful information and allows GeoMM to obtain much better accuracy.

(5) Regression with proposed factorization. We pose BLI as a regression problem, as done in previous approaches, by employing the following loss function: $\|\mathbf{U}_t\mathbf{B}\mathbf{U}_s^\top\mathbf{X}_s - \mathbf{X}_t\|_F^2$. We observe that its performance is worse than the classification baseline ($\mathbf{W} \in \mathbb{R}^{d \times d}$). The classification setting directly models the similarity score via the loss function, and hence corresponds with inference more closely. This result further reinforces the observation made in the first ablation test.

To summarize, the proposed modeling choices are better than the alternatives compared in the ablation tests.

6.3 Cross-lingual Word Similarity

The results on the cross-lingual word similarity task using the SemEval 2017 dataset (Camacho-Collados et al., 2017) are shown in Table 4. We observe that GeoMM performs better than Procrustes, MSF, and the SemEval 2017 baseline NASARI (Camacho-Collados et al., 2016). It is also competitive with Luminoso_run2 (Speer and Lowry-Duda, 2017), the best reported system on this dataset. It should be noted that NASARI and luminoso_run2 use additional knowledge sources like BabelNet and ConceptNet.

Method	en-es	en-de	en-it
NASARI	0.64	0.60	0.65
Luminoso_run2	0.75	0.76	0.77
Procrustes	0.72	0.72	0.71
MSF	0.73	0.74	0.73
Joulin et al. (2018)	0.71	0.71	0.71
GeoMM	0.73	0.74	0.74

Table 4: Pearson correlation coefficient for the SemEval 2017 cross-lingual word similarity task.

Method	Accuracy (%)
Original English embeddings	76.66
Procrustes	76.66
MSF	76.59
GeoMM	75.21

Table 5: Results on the monolingual word analogy task.

6.4 Monolingual Word Analogy

Table 5 shows the results on the English monolingual analogy task after obtaining it→en mapping on the VecMap dataset (Mikolov et al., 2013a; Artetxe et al., 2016). We observe that there is no significant drop in the monolingual performance by the use of non-orthogonal mappings compared with monolingual embeddings as well as other bilingual embeddings (Procrustes and MSF).

7 Indirect Translation: Results and Analysis

In the previous sections, we have established the efficacy of our approach for a bilingual mapping problem when a bilingual dictionary between the source and target languages is available. We also showed that our proposed multilingual generalization (Section 3.4) performs well in this scenario. In this section, we explore if our multilingual generalization is beneficial when a bilingual dictionary is not available between the source and the target, in other words, *indirect translation*. For this evaluation, our algorithm learns a *single model* for various language pairs such that word embeddings of different languages are transformed to a common latent space.

7.1 Evaluation Task: One-hop Translation

We consider the BLI task from language L_{src} to language L_{tgt} in the absence of a bilingual lexicon between them. We, however, assume the availability of lexicons for L_{src}-L_{pvt} and L_{pvt}-L_{tgt}, where L_{pvt} is a *pivot* language.

Method	fr-it-pt	it-de-es	es-pt-fr	avg.
SL-unsup	74.1	86.4	84.6	81.7
Composition				
Procrustes	74.2	81.9	82.5	79.5
MSF	75.3	81.9	82.7	80.0
GeoMM	77.7	84.1	84.3	82.0
Pipeline				
Procrustes	72.5	61.6	79.9	71.3
MSF	75.9	64.5	82.5	74.3
GeoMM	75.9	62.5	81.7	73.4
GeoMM$_{\text{multi}}$	**80.1**	**86.8**	**85.6**	**84.2**

Table 6: Indirect translation: Precision@1 for BLI.

As baselines, we adapt any supervised bilingual approach (Procrustes, MSF, and the proposed GeoMM) to the one-hop translation setting by considering their following variants:

- **Composition** (cmp): Using the given bilingual approach, we learn the $L_{\text{src}} \rightarrow L_{\text{pvt}}$ and $L_{\text{pvt}} \rightarrow L_{\text{tgt}}$ transformations as $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. Given an embedding x from L_{src}, the corresponding embedding in L_{tgt} is obtained by a composition of the transformations, *i.e.,* $\mathbf{W}_2\mathbf{W}_1x$. This is equivalent to computing the similarity of L_{src} and L_{tgt} embeddings in the L_{pvt} embedding space. Recently, Smith et al. (2017a) explored this technique with the Procrustes algorithm.

- **Pipeline** (pip): Using the given bilingual approach, we learn the $L_{\text{src}} \rightarrow L_{\text{pvt}}$ and $L_{\text{pvt}} \rightarrow L_{\text{tgt}}$ transformations as $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. Given a word embedding x from L_{src}, we infer its translation embedding z in L_{pvt}. Then, the corresponding embedding of x in L_{tgt} is $\mathbf{W}_2z$.

As discussed in Section 3.4, our framework allows the flexibility to jointly learn the common latent space of multiple languages, given bilingual dictionaries of multiple language pairs. Our multilingual approach, GeoMM$_{\text{multi}}$, views this setting as a graph with three nodes $\{L_{\text{src}}, L_{\text{tgt}}, L_{\text{pvt}}\}$ and two edges $\{L_{\text{src}}\text{-}L_{\text{pvt}}, L_{\text{pvt}}\text{-}L_{\text{tgt}}\}$ (dictionaries).

7.2 Experimental Settings

We experiment with the following one-hop translation cases: (a) fr-it-pt, (b) it-de-es, and (c)

Method	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-it	it-en	avg.
RCSLS	**84.1**	86.3	**83.3**	84.1	**79.1**	76.3	**57.9**	67.2	45.9	**46.4**	45.1	38.3	66.2
GeoMM	81.9	85.5	82.1	84.2	74.9	**76.7**	52.8	67.6	**49.1**	45.3	48.3	41.2	65.8
GeoMM$_{\text{semi}}$	82.7	**86.7**	82.8	**84.9**	76.4	**76.7**	53.2	**68.2**	48.5	46.1	**50.0**	**42.6**	**66.6**

Table 7: Comparison of GeoMM and GeoMM$_{\text{semi}}$ with RCSLS (Joulin et al., 2018). Precision@1 for BLI is reported. The results of RCSLS are reported in the original paper. The results of language pairs en-it and it-en are on the VecMap dataset, while others are on the MUSE dataset.

es-pt-fr (read the triplets as L_{src}-L_{pvt} -L_{tgt}). The training/test dictionaries and the word embeddings are from the MUSE dataset. In order to minimize direct transfer of information from L_{src} to L_{tgt}, we generate L_{src}-L_{pvt} and L_{pvt}-L_{tgt} training dictionaries such that they do not have any L_{pvt} word in common. The training dictionaries have the same size as the L_{src}-L_{pvt} and L_{pvt}-L_{tgt} dictionaries provided in the MUSE dataset while the test dictionaries have $1{,}500$ entries.

7.3 Results and Analysis

Table 6 shows the results of the one-hop translation experiments. We observe that GeoMM$_{\text{multi}}$ outperforms pivoting methods (cmp and pip) built on top of MSF and Procrustes for all language pairs. It should be noted that pivoting may lead to cascading of errors in the solution, whereas learning a common embedding space jointly mitigates this disadvantage. This is reaffirmed by our observation that GeoMM$_{\text{multi}}$ performs significantly better than GeoMM (cmp) and GeoMM (pip).

Since unsupervised methods have been shown to be competitive with supervised methods, they can be an alternative to pivoting. Indeed, we observe that the unsupervised method SL-unsup is better than the pivoting methods, although it used no bilingual dictionaries. On the other hand, GeoMM$_{\text{multi}}$ is better than the unsupervised methods too. It should be noted that the unsupervised methods use much larger vocabulary than GeoMM$_{\text{multi}}$ during the training stage.

We also experimented with scenarios where some words from L_{pvt} occur in both L_{src}-L_{pvt} and L_{pvt}-L_{tgt} training dictionaries. In these cases too, we observed that GeoMM$_{\text{multi}}$ perform better than other methods. We have not included these results because of space constraints.

8 Semi-supervised GeoMM

In this section, we discuss an extension of GeoMM, which benefits from unlabeled data. For the bilingual mapping problem, unlabeled data is available in the form of vocabulary lists for both the source and target languages. Existing unsupervised and semi-supervised techniques (Artetxe et al., 2017, 2018b; Joulin et al., 2018; Hoshen and Wolf, 2018) have an iterative refinement procedure that employs the vocabulary lists to augment the dictionary with positive or negative mappings.

Method	en-it	en-de	en-fi	en-es	avg.
GeoMM	48.3	49.3	36.1	39.3	43.3
GeoMM$_{\text{semi}}$	**50.0**	**51.3**	**36.2**	**39.7**	**44.3**

Table 8: Precision@1 for BLI on the VecMap dataset.

Given a seed bilingual dictionary, we implement a bootstrapping procedure that iterates over the following two steps until convergence:

1. Learn the GeoMM model by solving the proposed formulation (1) with the current bilingual dictionary.

2. Compute a new bilingual dictionary from the vocabulary lists, using the (current) GeoMM model for retrieval. The seed dictionary along with this new dictionary is used in the next iteration.

In order to keep the computational cost low, we restrict the vocabulary list to k most frequent words for both the languages (Artetxe et al., 2018b; Hoshen and Wolf, 2018). In addition, we perform bidirectional dictionary induction (Artetxe et al., 2018b; Hoshen and Wolf, 2018). We track the model's performance on a validation set to avoid overfitting and use it as a criterion for convergence of the bootstrap procedure.

We evaluate the proposed semi-supervised GeoMM algorithm (referred to as **GeoMM$_{\text{semi}}$**) on the bilingual lexicon induction task on MUSE and VecMap datasets. The bilingual dictionary for training is split $80/20$ into the seed dictionary

and the validation set. We set $k = 25{,}000$, which works well in practice.

We compare GeoMM$_{\text{semi}}$ with RCSLS, a recently proposed state-of-the-art semi-supervised algorithm by Joulin et al. (2018). RCSLS directly optimizes the CSLS similarity score (Conneau et al., 2018), which is used during retrieval stage for GeoMM, among other algorithms. On the other hand, GeoMM$_{\text{semi}}$ optimizes a simpler classification-based square loss function (see Section 3.3). In addition to the training dictionary, RCSLS uses the full vocabulary list of the source and target languages (200,000 words each) during training.

The results are reported in Table 7. We observe that the overall performance of GeoMM$_{\text{semi}}$ is slightly better than RCSLS. In addition, our supervised approach GeoMM performs slightly worse than RCSLS, although it does not have the advantage of learning from unlabeled data, as is the case for RCSLS and GeoMM$_{\text{semi}}$. We also notice that GeoMM$_{\text{semi}}$ improves on GeoMM in almost all language pairs.

We also evaluate GeoMM$_{\text{semi}}$ on the VecMap dataset. The results are reported in Table 8. To the best of our knowledge, GeoMM$_{\text{semi}}$ obtains state-of-the-art results on the VecMap dataset.

9 Conclusion and Future Work

In this work, we develop a framework for learning multilingual word embeddings by aligning the embeddings for various languages in a common space and inducing a Mahalanobis similarity metric in the common space. We view the translation of embeddings from one language to another as a series of geometrical transformations and jointly learn the language-specific orthogonal rotations and the symmetric positive definite matrix representing the Mahalanobis metric. Learning such transformations can also be viewed as learning a suitable common latent space for multiple languages. We formulate the problem in the Riemannian optimization framework, which models the above transformations efficiently.

We evaluate our bilingual and multilingual algorithms on the bilingual lexicon induction and the cross-lingual word similarity tasks. The results show that our algorithm outperforms existing approaches on multiple datasets. In addition, we demonstrate the efficacy of our multilingual algorithm in a one-hop translation setting for bilingual lexicon induction, in which a direct dictionary between the source and target languages is not available. The semi-supervised extension of our algorithm shows that our framework can leverage unlabeled data to obtain further improvements. Our analysis shows that the combination of the proposed transformations, inference in the induced latent space, and modeling the problem in classification setting allows the proposed approach to achieve state-of-the-art performance.

In future, an unsupervised extension to our approach can be explored. Optimizing the CSLS loss function (Joulin et al., 2018) within our framework can be investigated to address the hubness problem. We plan to work on downstream applications like text classification, machine translation, *etc.*, which may potentially benefit from the proposed latent space representation of multiple languages by sharing annotated resources across languages.

References

Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. 2008. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. Technical report, arXiv preprint arXiv:1602.01925.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 789–798. https://github.com.artetxem/vecmap.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*.

Silvère Bonnabel and Rodolphe Sepulchre. 2010. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070.

Nicolas Boumal, Bamdev Mishra, Pierre-Antoine Absil, and Rodolphe Sepulchre. 2014. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(Apr):1455–1459.

José Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation*.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1853–1861.

Xilun Chen and Claire Cardie. 2018. Unsupervise multilingual word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations*. https://github.com/facebookresearch/MUSE.

Georgiana Dinu and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Workshop track of International Conference on Learning Representations*.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. https://github.com/yeraidam/meemi.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 894–904.

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

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

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the International Conference on Machine Learning*, pages 748–756.

John C. Gower. 1975. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with Wasserstein Procrustes. Technical report, arXiv preprint arXiv:1805.11222.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine

translation for extremely low resource languages. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mehrtash Harandi, Mathieu Salzmann, and Richard Hartley. 2017. Joint dimensionality reduction and metric learning: A geometric take. In *Proceedings of the International Conference on Machine Learning*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 58–68.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 469–478.

Kejun Huang, Matt Gardner, Evangelos E. Papalexakis, Christos Faloutsos, Nikos D. Sidiropoulos, Tom M. Mitchell, Partha Pratim Talukdar, and Xiao Fu. 2015. Translation invariant word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1084–1088.

Wen Huang, Pierre-Antoine Absil, Kyle A. Gallivan, and Paul Hand. 2016. ROPTLIB: An object-oriented C++ library for optimization on Riemannian manifolds. Technical report, FSU16-14.v2, Florida State University.

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

Y. Kementchedjhieva, S. Ruder, R. Cotterell, and A. Søgaard. 2018. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics: Technical Papers*, pages 1459–1474.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1959–1970.

John M. Lee. 2003. *Introduction to Smooth Manifolds*, second edition, Springer-Verlag, New York.

Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. 2011. Linear regression under fixed-rank constraints: A Riemannian approach. In *Proceedings of the International Conference on Machine Learning*, pages 545–552.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. Technical report, arXiv preprint arXiv:1301.3781.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. Technical report, arXiv preprint arXiv:1309.4168.

Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. 2014. Fixed-rank matrix factorizations and Riemannian low-rank optimization. *Computational Statistics*, 29(3):591–621.

Hiroyuki Sato and Toshihiro Iwai. 2013. A new, globally convergent Riemannian conjugate gradient method. *Optimization: A Journal of Mathematical Programming and Operations Research*, 64(4):1011–1031.

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. 2017a. Aligning the fastText vectors of 78 languages. URL: `https://github.com/Babylonpartners/fastText_multilingual`.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017b. Offline bilingual word vectors, orthogonal

transformations and the inverted softmax. In *Proceedings of the International Conference on Learning Representations*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 778–788.

Robert Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 Task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluations*.

James Townsend, Niklas Koep, and Sebastian Weichwald. 2016. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5. URL: `https://pymanopt.github.io`.

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

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 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 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945.

Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 430–440.