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

Pratik Jawanpuria[*] PRATIK.JAWANPURIA@MICROSOFT.COM
Arjun Balgovind[†] ARJUNBALGOVIND@GMAIL.COM
Anoop Kunchukuttan[*] ANKUNCHU@MICROSOFT.COM
Bamdev Mishra[*] BAMDEVM@MICROSOFT.COM

Abstract

We propose a novel geometric approach for learning bilingual mappings given monolingual embeddings and a bilingual dictionary. Our approach decouples learning the transformation from the source language to the target language into (a) learning rotations for language-specific embeddings to align them to a common space, and (b) learning a similarity metric in the common space to model similarities between the embeddings. We model the bilingual mapping problem as an optimization problem on smooth Riemannian manifolds. We show that our approach outperforms previous approaches on the bilingual lexicon induction and cross-lingual word similarity tasks. We also generalize our framework to represent multiple languages in a common latent space. In particular, the latent space representations for several languages are learned jointly, given bilingual dictionaries for multiple language pairs. We illustrate the effectiveness of joint learning for multiple languages in zero-shot word translation setting.

1. Introduction

Bilingual words embeddings are a useful tool in natural language processing that has attracted a lot of interest lately. This interest stems from their fundamental property that similar concepts/words across different languages are mapped close to each other in a common space. Such embeddings are useful for different multilingual applications like machine translation (Gu et al., 2018), building bilingual dictionaries (Mikolov et al., 2013b), mining parallel corpora (Conneau et al., 2018), *etc.* They are also useful for sharing annotated corpora/linguistic resources across languages via transfer learning and joint learning for various tasks like text classification (Klementiev et al., 2012), sentiment analysis (Zhou et al., 2015), dependency parsing (Ammar et al., 2016), *etc.*

Mikolov et al. (2013b) showed encouraging results with a simple approach to learn bilingual embeddings. They empirically show that a linear transformation of embeddings from one language to another preserves the geometric arrangement of word embeddings. Specifically, let x represents the embedding of a word w_x in the *source* language, and let $\mathbf{W}$ be the (learned) linear transformation from the source language to the *target* language. A

[*] Microsoft, India.
[†] IIT Madras. This work was carried out during the author's internship at Microsoft, India.

suitable approximation of the embeddings of the translation of the word w_x in the target language is $\mathbf{W}x$ (Mikolov et al., 2013b). In a supervised setting, $\mathbf{W}$ is learned given a small bilingual dictionary and their corresponding monolingual embeddings as training data. Subsequently, many refinements have been proposed to this *offline* bilingual mapping framework (Xing et al., 2015; Smith et al., 2017a; Conneau et al., 2018; Artetxe et al., 2016, 2017, 2018a,b).

In this work, we propose a novel geometric approach for the *offline* mapping problem. We learn language specific orthogonal transformations to align the source and target language embeddings, thereby mapping them to a common space. We leverage this common space to learn a similarity metric, also known as the Mahalanobis metric, between the word embeddings. This metric accounts for the feature correlation and is learned *globally* using all the training data. It primarily refines the notion of distance between a pair of embeddings from two different languages, based on the available evidence*, i.e.,* the training data. Overall, our model decouples the transformation matrix $\mathbf{W}$ from the source language to the target language into the language specific orthogonal transformations as well as the similarity metric 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. The matrix $\mathbf{B}$ represents the Mahalanobis metric and it generalizes the notion of cosine similarity in the Euclidean space.

A key feature of our framework is that it naturally generalizes to multilingual settings*, i.e.,* the ability to represent embeddings from multiple languages in a single vector space. For each language L_i, the corresponding transformation matrix $\mathbf{U}_i$ aligns its embeddings to the common space. The Mahalanobis metric ($\mathbf{B}$) learned in this common space captures feature correlation information of the word embeddings across all the given languages. Given bilingual dictionaries for multiple language pairs, our approach learns the transformation matrices as well as the similarity metric jointly.

The proposed optimization problem involves orthogonal constraints on language specific transformations ($\mathbf{U}_i$) as well as the 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 problem over the manifolds. In the process, it is able to exploit the geometry of the manifolds and the symmetries involved in them. We propose to solve the resulting optimization problem efficiently with the Riemannian conjugate-gradient algorithm (Absil et al., 2008).

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

- We propose a novel approach for learning bilingual embeddings by learning representations in a common latent space. This involves aligning word embeddings of the language pair via suitable orthogonal transformations and inducing a similarity metric to refine the notion of distance between the word embeddings. The model naturally generalizes to learn multilingual languages.

- We show that our approach outperforms state-of-the-art method on the bilingual lexicon induction task. We also outperform existing methods on the cross-lingual word similarity task.
- An ablation analysis of various components of our model shows that the following contribute to the improved performance of our model: (a) ability to learn a similarity metric in the latent space, and (b) formulating the problem in the classification setting, which is closely related to the inference stage.
- In a multilingual setting, we learn a single model given bilingual dictionaries of multiple language pairs. We map multiple languages into a single vector space by learning the characteristics common across languages (via the Mahalanobis metric) as well as language specific attributes (the orthogonal transformations). We evaluate our multilingual model in *zero-shot word translation* problem: language pairs do not have a bilingual dictionary between them but have bilingual dictionary with some pivot language.

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 frameworks and analyzes the results. Section 7 presents experiments on indirect translation using the multilingual extensions to our models. Section 8 concludes the paper.

2. Related Work

In this section, we summarize previous work on supervised as well as unsupervised learning of bilingual learning. We also report previous work on multilingual embeddings.

2.1 Supervised Learning of Bilingual Embeddings

Mikolov et al. (2013b) showed that a linear transformation from embeddings of one language to another gives promising results. They proposed to solve the following *regression problem* to learn the transformation:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times d}} \|\mathbf{W}\mathbf{X}_s - \mathbf{X}_t\|_F^2, \tag{1}$$

where the matrices $\mathbf{X}_s$ and $\mathbf{X}_t$ correspond to source and target languages word embeddings, respectively. Each column in these matrices represents a d dimensional word embedding. The linear transformation matrix is denoted by $\mathbf{W}$. This approach is categorized as an *offline* method since the monolingual and bilingual embeddings are learned separately. In contrast *online* approaches directly learn a bilingual embedding from parallel corpora (Hermann and Blunsom, 2014), 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.

Existing works have suggested improvements upon the framework proposed by Mikolov et al. (2013b). One popular modification in (1) is to constraint $\mathbf{W}$ to be an orthogonal matrix, *i.e.,* the learned $\mathbf{W}$ satisfies the constraint $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_d$, where $\mathbf{I}_d$ is the identity matrix of dimension d. Several works (Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017a) study the following formulation for learning the bilingual mappings of word embeddings:

$$\min_{\mathbf{W}\in\mathbb{O}^d} \|\mathbf{W}\mathbf{X}_s - \mathbf{X}_t\|_F^2, \tag{2}$$

where $\mathbb{O}^d$ denotes the space of d dimensional orthogonal matrices. Various motivations for the orthogonality constraint have been suggested: (1) preserving monolingual inference (Artetxe et al., 2016), (2) length normalization of the embeddings (Xing et al., 2015), and (3) ensuring mappings are reversible (Smith et al., 2017a). It should be noted that (2) is the well-known *orthogonal Procrustes problem* and its solution has the following form (Schönemann, 1966): $\mathbf{W} = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices having closed-form expressions (in terms of $\mathbf{X}_s$ and $\mathbf{X}_t$). It can also be shown that if the word embeddings in $\mathbf{X}_s$ and $\mathbf{X}_t$ are unit normalized, optimizing (2) is equivalent to maximizing the cosine similarity between $\mathbf{V}^\top\mathbf{X}_s$ and $\mathbf{U}^\top\mathbf{X}_t$ (Smith et al., 2017a). In addition, Artetxe et al. (2016, 2018a) have suggested pre-processing and post-processing transformations to the data that have empirically yielded better solutions over the vanilla Procrustes solution.

Existing works have also explored other loss functions as well in the bilingual mapping problem. In order to rank the correct translation of a word embedding higher than other possible incorrect translations, Lazaridou et al. (2015) explored a margin-based ranking loss function (Weston et al., 2010, 2011) without the orthogonality constraint on $\mathbf{W}$. Recently, Joulin et al. (2018) propose to optimize the cross-lingual similarity scaling score (Conneau et al., 2018) for learning the transformation matrix $\mathbf{W}$. In their optimization problem, they relax the non-convex $\mathbf{W} \in \mathbb{O}^d$ constraint with the convex constraints: the unit spectral norm (or the unit Frobenius norm) constraint on $\mathbf{W}$. They observe a decrease in bilingual mapping performance if they enforce $\mathbf{W} \in \mathbb{O}^d$ in their solution.

Another approach, Canonical Correlation Analysis (CCA), also learns linear projections from the source and target languages to a common space such that correlations between the embeddings projected to this space are maximized (Faruqui and Dyer, 2014). Previous works (Artetxe et al., 2016, 2018a) show that the Procrustes solution based approaches perform better than the CCA based techniques.

We propose a novel factorization for the transformation matrix $\mathbf{W}$ and learn a common latent space for the word embeddings of both source and target languages. In addition, we formulate the task of learning bilingual mappings of word embeddings as a classification problem: predicting if the given word pairs are translations of each other or not. The classification setting allows the learning problem to adapt closely to the inference stage.

2.2 Retrieval Method for Bilingual Lexicon Induction

The simplest method for obtain the translation of a word is to perform a nearest neighbour search in the projected space among the target language words. However, the nearest neighbour (NN) search in high dimensional space encounters the *hubness* problem, specifically, there exist *hub* points in the space that are closer to many points. To mitigate this problem, alternative search procedures which penalize hubs have been proposed: inverted rank (Dinu and Baroni, 2015), inverted soft-max (Smith et al., 2017a), and cross-domain similarity local scaling (CSLS) (Conneau et al., 2018). The CSLS search, in particular, has show significant improvements over the NN search over a wide range of datasets and training methods (Conneau et al., 2018; Artetxe et al., 2018b). Hence, we employ the CSLS search based retrieval method in our experiments.

2.3 Multilingual Embeddings

Most offline methods have focussed on the bilingual scenario only and the multilingual setting has received less attention. Multilingual embeddings make zero-shot word translation possible. Existing works demonstrate the efficacy of CCA and regression models in bilingual scenarios only. A simple method to adapting bilingual embeddings for representing embeddings of multiple languages in a common vector space is to designate one of the languages as a *pivot* language. Bilingual mappings are learned *independently* from all other languages to the pivot language. Using this approach, previous works (Ammar et al., 2016; Smith et al., 2017b) have proposed building multilingual embeddings from bilingual embeddings.

We propose to align the embeddings of all the languages into a common latent space. In particular, we learn language specific feature transformation and a common similarity metric $\mathbf{B}$ *jointly*. This allows our model to also leverage the commonality across the word embeddings of various languages.

Existing works have also proposed online methods that can be trained to learn a common embedding space across languages. However, these require more resources like parallel sentence corpora (Huang et al., 2015; Duong et al., 2017).

2.4 Unsupervised Learning of Bilingual Embeddings

As mentioned above, multilingual embeddings provide a solution to map embedding spaces between languages for which we do not possess a bilingual dictionary. An alternative solution is unsupervised learning of the mapping function*, i.e.,* given only the monolingual embeddings of the two languages. The simplest of these is a semi-supervised solution which uses a small seed bilingual dictionary and an iterative learning process (Artetxe et al., 2017). First, a mapping function is learnt using the seed dictionary using a supervised solution. The seed dictionary is then augmented by finding high-confidence translation pairs. These two steps are repeated iteratively until convergence. The small seed dictionary is cheap to construct: (i) containing 25-50 word pairs, and/or (ii) containing numerals or identically spelt words which can be automatically mined. Artetxe et al. (2018b) and Hoshen and Wolf (2018) have proposed initialization methods that do away with the need for a seed dictionary. Recently, Zhang et al. (2017b); Grave et al. (2018) proposed aligning the the source and target language word embeddings by optimizing the earth mover's distance or the Wasserstein distance. Unsupervised methods based on adversarial training objectives have also been proposed (Barone, 2016; Zhang et al., 2017a; Conneau et al., 2018). They enforce similar source and target language words to have similar embeddings.

3. Learning Latent Space Representations

In this section, we describe the proposed framework to learn the multilingual mapping of word embeddings. We first proceed with bilingual mapping problem and later generalize our framework to multilingual setting.

3.1 Geometry-aware Factorization of the Transformation Matrix

We propose to transform the word embeddings from source language (S) and target language (T) to a common space in which the semantic similarity of words from different languages may be better learned. To this end, we propose to learn orthogonal transformations (rotations) $\mathbf{U}_s \in \mathbb{O}^d$ and $\mathbf{U}_t \in \mathbb{O}^d$ for source and target languages embeddings, respectively, where $\mathbb{O}^d$ represents the space of d dimensional orthogonal matrices. Hence, a word embedding x corresponding to word w_x in language S is transformed to $\psi_s(x) = \mathbf{U}_s^\top x$. Similarly, a word embedding z corresponding to word w_z in language T is transformed to $\psi_t(z) = \mathbf{U}_t^\top z$. The aim of learning these orthogonal transformations is to map both the source and target languages to a common space in which we can learn a suitable similarity metric, as discussed below.

We further propose to learn a Mahalanobis metric $\mathbf{B}$ to *refine* the notion of similarity[1] between the two transformed embeddings $\psi_s(x)$ and $\psi_t(z)$. Since $\mathbf{B}$ is a metric in $\mathbb{R}^d$ space, it is constrained to be a $d \times d$ positive definite matrix $\mathbf{B}$*, i.e.,* $\mathbf{B} \succ \mathbf{0}$. $\mathbf{B}$ accounts for the feature correlation information from the training data which allows a more informative comparison of language embeddings than the vanilla cosine similarity. In fact, Mahalanobis similarity reduces to cosine similarity when $\mathbf{B} = \mathbf{I}$*, i.e.,* when the features are uncorrelated, which may be a strong assumption in real world applications. 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 linear transformation for language S to T is expressed as $\mathbf{W}_{ts} = \mathbf{U}_t \mathbf{B} \mathbf{U}_s^\top$. For a given embedding x in language S, its corresponding embedding in language T is given by $\mathbf{W}_{ts} x$.

The proposed factorization of the (bilingual) transformation matrix as $\mathbf{W} = \mathbf{U}\mathbf{B}\mathbf{V}^\top$, 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). Recent works have explored it in low-rank matrix completion (Mishra et al., 2014) and low-rank metric learning applications (Harandi et al., 2017). Polar factorization has similarity with the singular value decomposition (SVD) of a matrix. The main 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).

Latent Space Interpretation of our Model

In the following, we present an equivalent latent space interpretation of our proposed model in which the Mahalanobis metric can be viewed as another linear feature transformation $\phi : \mathbb{R}^d \to \mathbb{R}^d$ of the embeddings. This mapping may be defined as $\phi(w) = \mathbf{B}^{\frac{1}{2}} w$, where $\mathbf{B} \succ \mathbf{0}$. Since $\mathbf{B}$ is a symmetric positive definite matrix, $\mathbf{B}^{\frac{1}{2}}$ is well defined and unique.

Our model may now be represented as learning a suitable latent feature space as follows. The source and target language embeddings are linearly transformed as $x \mapsto \phi(\psi_s(x))$ and $z \mapsto \phi(\psi_t(z))$, respectively. The linear 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 $\mathbf{U}_t$ as linear operators corresponding to the linear functions $\phi(\cdot)$, $\psi_s(\cdot)$, and $\psi_t(\cdot)$,

[1] Mahalanobis metric generalizes the notion of cosine similarity in the Euclidean space. For given two unit normalized vectors $x_1 \in \mathbb{R}^d$ and $x_2 \in \mathbb{R}^d$, their cosine similarity is given by $\mathrm{sim}_{\mathbf{I}}(x_1, x_2) = x_1^\top \mathbf{I} x_2 = x_1^\top x_2$, where $\mathbf{I}$ is the d-dimensional identity matrix. However, if this space is endowed with metric $\mathbf{B}$, their cosine similarity is computed as $\mathrm{sim}_{\mathbf{B}}(x_1, x_2) = x_1^\top \mathbf{B} x_2$.

respectively. Since the matrix $\mathbf{B}$ is embedded implicitly in this latent feature space, we employ the usual cosine similarity measure, computed as $\phi(\psi_t(z))^\top \phi(\psi_s(x)) = z^\top \mathbf{U}_t \mathbf{B} \mathbf{U}_s^\top x$ $(= h_{st}(x, z))$. We can observe that the similarity function in both the interpretations of our model is the same $(h_{st}(x, z))$.

3.2 Proposed Formulation for the Bilingual Mapping Problem

We learn the orthogonal transformations $\mathbf{U}_s \in \mathbb{O}^d$ and $\mathbf{U}_t \in \mathbb{O}^d$ as well as the metric $\mathbf{B} \succ \mathbf{0}$ by minimizing a loss function in the supervised setting. In bilingual mapping problem, we assume that a small bilingual dictionary (of size n) is available as the training data. Let $\mathbf{X}_s \in \mathbb{R}^{d \times n_s}$ denote the embeddings of the dictionary words from the source language, where d is the dimensionality of the embeddings and n_s is the number of unique words from the source language present in the dictionary. Similarly, let $\mathbf{X}_t \in \mathbb{R}^{d \times n_t}$ denote the embeddings of the dictionary words from the target language, where n_t is the number of unique words from the target language present in the dictionary.

We propose to model the bilingual word embedding mapping problem as a binary classification problem. Consider the i-th dictionary word embedding x from the source language and the j-th dictionary word embedding z from the target language. If the words corresponding to x and z map correctly 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 $h_{st}(x, z)$. Since the prediction function $h_{st}(x, z)$ is a cosine similarity meaning between a pair of embeddings, it is invariant of the direction of mapping: whether it is source to target language or target language to source language $(h_{st}(x, z) = h_{ts}(z, x))$. Hence, our model learns bidirectional mapping (source to target language and vice-versa) via a single optimization problem. More concretely, the linear transformation for language T to S is $\mathbf{W}_{st} = \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top$, *i.e.,* $\mathbf{W}_{st} = \mathbf{W}_{ts}^\top$. We create a binary $0-1$ label matrix $\mathbf{Y}_{st} \in \{0, 1\}^{n_s \times n_t}$ whose (i, j)-th entry corresponds to the correctness in mapping of i-th word from the source language and j-th word from the target language. Let Ω be the set of row-column indices corresponding to entry value 1 in the matrix $\mathbf{Y}_{st}$. Then, $|\Omega|$ is equal to the size of given bilingual dictionary. The overall optimization problem for learning bilingual mappings is as follows:

$$\min_{\mathbf{U}_s \in \mathbb{O}^d, \mathbf{U}_t \in \mathbb{O}^d, \mathbf{B} \succ \mathbf{0}} \lambda \|\mathbf{B}\|_F^2 + \ell(\mathbf{X}_s^\top \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top \mathbf{X}_t, \mathbf{Y}_{st}), \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\lambda > 0$ is the regularization parameter, and ℓ is a suitable loss function. In particular, we employed the square loss[2] since it is smooth and relatively easier to optimize. Our optimization problem with the square loss is as follows:

$$\min_{\mathbf{U}_s \in \mathbb{O}^d, \mathbf{U}_t \in \mathbb{O}^d, \mathbf{B} \succ \mathbf{0}} \lambda \|\mathbf{B}\|_F^2 + \|\mathbf{X}_s^\top \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top \mathbf{X}_t - \mathbf{Y}_{st}\|_F^2. \tag{4}$$

Though the loss term in the proposed optimization problem (4) is represented as a function of $n_s \times n_t$ matrix, its computation complexity is a linear function of $|\Omega|$. This is because the label matrix $\mathbf{Y}_{st}$ is binary valued, which can be exploited to rewrite the loss term in

[2] Other similarity score based loss functions may also be employed in our framework.

(4) as follows:

$$
\begin{aligned}
\|\mathbf{X}_s^\top \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top \mathbf{X}_t - \mathbf{Y}_{st}\|_F^2 &= \|\mathbf{X}_s^\top \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top \mathbf{X}_t\|_F^2 + \|\mathbf{Y}_{st}\|_F^2 - 2\mathrm{Tr}(\mathbf{Y}_{st}^\top \mathbf{X}_s^\top \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top \mathbf{X}_t), \\
&= \mathrm{Tr}\big(\mathbf{U}_t \mathbf{B} \mathbf{U}_s^\top (\mathbf{X}_s \mathbf{X}_s^\top) \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top (\mathbf{X}_t \mathbf{X}_t^\top)\big) + |\Omega| \\
&\quad - 2 \sum_{\{(i,j):(i,j)\in\Omega\}} x_{si}^\top \mathbf{U}_s \mathbf{B} \mathbf{U}_t^\top x_{tj},
\end{aligned}
\tag{5}
$$

where x_{si} represents the i-th column in $\mathbf{X}_s$, x_{tj} represents the j-th column in $\mathbf{X}_t$, and $\mathrm{Tr}(\cdot)$ denotes the trace of a square matrix. The complexity of computing the first and third term in (5) is $O(d^3 + n_s d^2 + n_t d^2)$ and $O(|\Omega| d + n_s d^2 + n_t d^2)$, respectively. Similarly, the computation cost of the gradient of the objective function in problem (4) with respect to the optimization variables $\mathbf{U}_s$, $\mathbf{U}_t$, and $\mathbf{B}$ is also a linear function of $|\Omega|$. Using the above trick, the same computational complexity may be obtained with any binary-valued label matrix*, i.e.,* $\mathbf{Y}_{st}$ has entries from the set $\{a, b\}$, where a and b are real numbers. Hence, our framework can efficiently leverage information from all the negative samples. This is in contrast with the work of Lazaridou et al. (2015), which selects a small subset of negative samples in order to avoid high computational complexity.

In the next section, we discuss a generalization of our approach for learning linear mappings in multilingual settings.

3.3 Generalization to Multilingual Setting

A common setting in real world applications is that the bilingual dictionary between source S and target T languages is not available but bilingual dictionaries of S and T with respect to a pivot language P is available. In this section, we propose a unified framework for learning multiple mappings when bilingual dictionaries are available for several language pairs.

We formalize this generalized setting as an undirected, connected graph $G(V, E)$, where each node represents a language and the edges between nodes represents the availability of a bilingual dictionary between the pair of language. V represents the sets of nodes (languages) and E represents the set of edges (availability of a bilingual dictionary between a pair of languages). A connected graph implies that one may *traverse* from any language to any other language via intermediate pivot language(s). In our context, this implies that word embeddings from one language can be translated to another language with the help of pivot language(s). Equivalently, we align the word embeddings of all the languages, and learn a common latent space for these languages.

We propose to learn a (single) orthogonal transformation $\mathbf{U}_i \in \mathbb{O}^d$ for every language L_i and a common Mahalanobis metric $\mathbf{B} \succ \mathbf{0}$. It should be noted that the transformation $\mathbf{U}_i$ for language L_i is common for all the bilingual mapping problems in this graph associated with language L_i. After the training stage is over, the linear transformation of translating an embedding x from language L_i to language L_j is given by $\mathbf{W}_{ji} = \mathbf{U}_j \mathbf{B} \mathbf{U}_i^\top$. It should be noted that we are able to obtain linear transformation even for language pairs that do have have any bilingual dictionary during the training stage.

Let $\mathbf{X}_i^j \in \mathbb{R}^{d\times m}$ be[3] the embeddings of the dictionary words of language i in the dictionary corresponding to edge $e_{ij} \in E$. Let $\mathbf{Y}_{ij} \in \{0,1\}^{m\times m}$ binary label matrix corresponding to the dictionary between languages i and j. The proposed optimization problem for learning multilingual mappings is as follows:

$$\min_{\mathbf{U}_i \in \mathbb{O}^d \forall i, \mathbf{B} \succ \mathbf{0}} \lambda \|\mathbf{B}\|_F^2 + \sum_{e_{ij}\in E} \|(\mathbf{X}_i^j)^\top \mathbf{U}_i \mathbf{B} \mathbf{U}_j^\top \mathbf{X}_j^i - \mathbf{Y}_{ij}\|_F^2. \tag{6}$$

We term our bilingual and multilingual mapping approach as **Geo**metry aware **M**ultilingual **M**apping (GeoMM). In the next section, we discuss the optimization algorithm for solving the bilingual mapping problem (4) as well as its generalization to the multilingual setting (6).

4. Optimization Algorithm

The geometric constraints $\mathbf{U}_s \in \mathbb{O}^d, \mathbf{U}_t \in \mathbb{O}^d$ and $\mathbf{B} \succ \mathbf{0}$ employed in the proposed problems (4) and (6) have been studied as smooth Riemannian manifolds, which are well explored topological spaces (Lee, 2003). The orthogonal matrice $\mathbf{U}_i$ 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 provides several tools to systematically optimize such problems (Absil et al., 2008). Overall, the Riemannian optimization framework generalizes several first- and second- order Euclidean algorithms (such as the conjugate-gradients and the trust-regions) to manifolds (Edelman et al., 1998; Absil et al., 2008; Journée et al., 2010; Sato and Iwai, 2013). We propose to optimize the problems (4) and (6) using the Riemannian conjugate gradient algorithm.

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 a wide range of batch/stochastic Riemannian optimization algorithms. We employ Python-based Pymanopt[4] in our experiments, where we only need to supply the objective function.

5. Experimental Settings

In this section, we describe the evaluation tasks, the datasets used, and other experimental details.

5.1 Evaluation Tasks

We evaluated our proposed approach on several tasks:

- To evaluate the quality of the bilingual mappings generated, we evaluate our method primarily for the bilingual lexicon induction (BLI) task, *i.e.,* word translation task and

[3] For notational convenience, we assume the unique words of every language in all their dictionaries to be same (m). However, our framework can easily accommodate dictionaries of different number of unique words in different dictionaries, as shown in our experiments.

[4] `https://pymanopt.github.io`.

compare Precision@1 with previously reported state-of-the-art results on the standard datasets (Dinu and Baroni, 2015; Artetxe et al., 2016; Conneau et al., 2018).

- In addition, we also evaluate our method on the cross-lingual word similarity task using the SemEval 2017 dataset.
- As a sanity check 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 a 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. We compare our multilingual embeddings with other approaches.

5.2 Datasets

For bilingual lexicon induction, we report results on two widely used, publicly available datasets.

- **VecMap**[5]: 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 to four languages: Italian (it), German (de), Finnish (fi) and Spanish (es). For each pair (en-it, en-de, en-fi, en-es), the dataset has predefined train and test bilingual dictionaries, pretrained monolingual embeddings (using word2vec). The details of these experimental settings can be found in Artetxe et al. (2018b).
- **MUSE**[6]: This dataset was originally made available by Conneau et al. (2018). It consists of pre-trained FastText (Bojanowski et al., 2017) monolingual embeddings trained on Wikipedia for many languages and predefined train and test bilingual dictionaries for various language pairs. Following Conneau et al. (2018), we report results[7] for English (en) to Spanish (es), French (fr), German (de), Russian (ru), Chinese (zh) and *vice versa*. The details of these experimental settings can be found in Conneau et al. (2018). In addition, we also experiment with English to Indian languages (Hindi, Tamil and Bengali) datasets using the same experimental settings.

5.3 Other Experimental Settings

- 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.
- In our evaluations, we use the latent space representations of word embeddings ($\mathbf{B}^{\frac{1}{2}}\mathbf{U}_i^\top x$) to compute distance/similarities between the embeddings.

[5] `https://github.com/artetxem/vecmap.`

[6] `https://github.com/facebookresearch/MUSE.`

[7] Conneau et al. (2018) also report English-Esparanto results, but this data is not available in the MUSE library.

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

We compare our proposed algorithm GeoMM with the best performing supervised methods. Additionally, we also compare with unsupervised methods as they have been shown to be competitive with supervised methods. We employ CSLS for retrieval (Conneau et al., 2018) unless otherwise mentioned. The following baselines are compared.

Supervised Baselines

1. Procrustes: the linear transformation for bilingual mapping is learned by solving the orthogonal Procrustes problem (Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017a). We employ the MUSE implementation.
2. MSF-ISF: the Multi-Step Framework proposed by Artetxe et al. (2018a). It employs the inverted softmax function (ISF) score for retrieval. We used the VecMap implementation.
3. MSF: the original work by Artetxe et al. (2018a) reported results using ISF for retrieval. We experimented with a variant of MSF-ISF, which uses CSLS for retrieval.
4. CSLS-Sp: a method that optimizes the CSLS score (Joulin et al., 2018). It relaxes the optimization problem by replacing the orthogonality constraint on the linear transformation with the spectral norm constraint.

Unsupervised Baselines

1. Adv-Refine: an unsupervised approach proposed by Conneau et al. (2018) that uses adversarial training with bilingual dictionary refinement in each iteration.
2. SL-unsup: state-of-the-art self-learning (SL) unsupervised model proposed by Artetxe et al. (2018b), which uses structural similarity of the embeddings to initialize the unsupervised training.
3. WS-Procrustes: an iterative unsupervised method (Grave et al., 2018) that solves the Procrustes problem in Wasserstein distance to learn a global alignment of embeddings in the two languages as well as the linear transformation.

6.1.1 VecMap Dataset

Table 1 reports the results on the VecMap dataset. We observe that our algorithm GeoMM obtains the best performance in each language pair, surpassing state-of-the-art results reported on this dataset. It should be noted that MSF (with CSLS inference) performs better than MSF-ISF (with ISF inference), which has been reported in literature.

Table 1: Precision@1 for bilingual lexicon induction on the VecMap dataset.

Method	**en-it**	**en-de**	**en-fi**	**en-es**	**avg.**
Supervised					
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
GeoMM	**48.3**	**49.3**	**36.1**	**39.3**	**43.3**
Unsupervised					
SL-unsup	48.1	48.2	32.6	37.3	41.6

Table 2: Precision@1 for bilingual lexicon induction on the MUSE dataset.

Method	**en-es**	**es-en**	**en-fr**	**fr-en**	**en-de**	**de-en**	**en-ru**	**ru-en**	**en-zh**	**zh-en**	**avg.**
Supervised											
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
CSLS-Sp	80.7	83.9	81.7	83.2	75.1	72.1	51.1	63.8	–	–	–
GeoMM	81.4	**85.5**	82.1	**84.1**	74.7	**76.7**	**51.3**	**67.6**	**49.1**	**45.3**	**69.8**
Unsupervised											
Adv-Refine	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	32.5	31.4	64.3
SL-unsup	82.3	84.7	82.3	83.6	75.1	74.3	49.2	65.6	0.0	0.0	59.7
WS-Procrustes	**82.8**	84.1	**82.6**	82.9	**75.4**	73.3	43.7	59.1	–	–	–

6.1.2 MUSE Dataset

Table 2 reports the results[8] on the MUSE dataset. We observe that our algorithm GeoMM obtains the best performance in most language pairs. GeoMM outperforms all the supervised algorithms, including CSLS-Sp which directly optimizes the CSLS score. It also outperforms the best unsupervised approaches on seven out of the ten language pairs, and has competitive results with the best algorithm for the remaining language pairs. It should be noted that the unsupervised algorithms have the entire vocabulary at their disposal during training, while the supervised results reported here do not use refinement to enhance the initial bilingual dictionary.

In addition to CSLS-Sp, Joulin et al. (2018) also propose the following variants of CSLS-Sp that employ the full vocabulary of both languages in addition to the given bilingual lexicon during training.

- CSLS-Sp-FV: it optimizes the CSLS score with respect to the full vocabulary (FV) of both languages as well as the given bilingual dictionary during the training stage.
- CSLS-Fr-FV: a variant of CSLS-Sp-FV that employs the Frobenius norm constraint instead of the spectral norm constraint.

It should be noted that other supervised methods that we evaluate (GeoMM, MSF, Procrustes) do not employ the full vocabulary during the training and optimize only on the

[8] Results for en-zh and zh-en have not been reported for CSLS-Sp by Joulin et al. (2018) and for WS-Procrustes by (Grave et al., 2018). SL-unsup gives 0.0 accuracy on en-zh and zh-en after multiple runs.

Table 3: Comparison of GeoMM with methods proposed by Joulin et al. (2018) which use additional information (Precision@1 for BLI).

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.**
CSLS-Fr-FV	**84.5**	**86.4**	**83.1**	**84.1**	**79.1**	75.9	**57.0**	67.1	44.6	41.9	–	–	–
CSLS-Sp-FV	83.0	84.9	82.7	**84.1**	78.2	75.8	56.4	66.4	44.4	**45.6**	45.3	37.9	65.4
GeoMM	81.4	85.5	82.1	**84.1**	74.7	**76.7**	51.3	**67.6**	**49.1**	45.3	**48.3**	**41.2**	**65.6**

Table 4: Precision@1 for BLI on Indian language ↔ English pairs on the MUSE dataset.

Method	**en-hi**	**hi-en**	**en-ta**	**ta-en**	**en-bn**	**bn-en**	**avg.**
Supervised							
Procrustes	33.3	42.8	15.1	20.8	15.8	24.6	25.4
MSF	40.4	**50.0**	**19.7**	25.3	21.1	30.6	31.2
GeoMM	**40.7**	49.5	19.5	**26.7**	**22.5**	**31.2**	**31.7**
Unsupervised							
SL-unsup	40.1	48.4	15.5	0.0	19.1	25.1	24.7

bilingual lexicon. However, for completeness, we also compare our algorithm with CSLS-Sp-FV and CSLS-Fr-FV. The results are reported[9] in Table 3. We observe that the overall performance of GeoMM is at par with both CSLS-Sp-FV and CSLS-Fr-FV. It should be noted that the en-it and it-en results are reported on the VecMap dataset, while the results for other language pairs are reported on the MUSE dataset.

6.1.3 Indian Languages

Most works on BLI have reported results on European languages. To understand the performance of BLI on diverse languages, we experiment with English to three Indic languages – Hindi (hi), Tamil (ta), and Bengali (bn) – and *vice versa*. Hindi and Bengali belong to the Indo-Aryan branch of the Indo-European language family, while Tamil belongs to the Dravidian language family. Table 4 reports the results of these experiments. We see that our algorithm GeoMM performs better than the best supervised as well as unsupervised methods.

In general, for all methods, we observe that the results for Indian languages are far inferior compared to the results for European languages. The degraded performance could be the result of multiple factors: (a) larger language divergence between English and Indian languages (compared to European languages), (b) the morphological richness of Indian languages (particularly agglutinative Dravidian languages), and (c) the quality of the monolingual embeddings (these have been trained on Wikipedia data, which is small for most for most Indian languages).

Table 5: Ablation test results: Precision@1 for BLI on the VecMap dataset.

Method	**en-it**	**en-de**	**en-fi**	**en-es**	**avg.**
GeoMM	**48.3**	**49.3**	**36.1**	**39.3**	**43.3**
Procrustes	44.9	46.5	33.5	35.1	40.0
Regression with proposed factorization	46.8	43.3	33.9	35.4	39.9
Classification with unconstrained **W**	45.4	47.9	35.4	37.5	41.5
No similarity metric	8.5	5.5	4.1	2.0	5.0
No language specific rotations	26.3	26.3	19.5	21.2	23.3

6.2 Ablation Tests

We study the impact of different components of our framework by varying one component at time. The results of these ablation tests are shown in Table 5 and discussed in the remainder of this section. We begin by analyzing if the choice of classification setting contributes to the improvements via the following ablation tests:

- **Regression with proposed factorization.** Instead of posing bilingual mapping as a classification task, we pose it as a regression task as done in most previous approaches to bilingual lexicon induction. The bilingual optimization problem can then be defined as:

$$\min_{\mathbf{U}_s \in \mathbb{O}^d, \mathbf{U}_t \in \mathbb{O}^d, \mathbf{B} \succ \mathbf{0}} \lambda \|\mathbf{B}\|_F^2 + \|\mathbf{U}_t \mathbf{B} \mathbf{U}_s^\top \mathbf{X}_s - \mathbf{X}_t\|_F^2. \tag{7}$$

  We observe that the classification setting is better than regression setting. We hypothesize that results for the classification setting are better since it is able to model the inference stage more closely.

- **Classification with unconstrained W.** We learn the linear transformation **W** directly in a classification setting without any constraints, *i.e.,* we solve the following optimization problem:

$$\min_{\mathbf{W} \in \mathbb{R}^d} \lambda \|\mathbf{W}\|_F^2 + \|\mathbf{X}_s^\top \mathbf{W}^\top \mathbf{X}_t - \mathbf{Y}_{st}\|_F^2. \tag{8}$$

  We observe that this setting shows better results than the Procrustes solution. This provides further evidence that the classification setting is more beneficial than the regression setting for bilingual mapping problem. It can be shown that optimization with unconstrained **W** yields the same objective value as the proposed problem (4), and the optimal $\hat{\mathbf{W}}$ of (8) may be obtained from the optimal solution $\{\hat{\mathbf{U}}, \hat{\mathbf{V}}, \hat{\mathbf{B}}\}$ of (4) as $\hat{\mathbf{W}} = \hat{\mathbf{U}}\hat{\mathbf{B}}\hat{\mathbf{V}}^\top$. The better results of GeoMM highlight importance of learning the proposed common latent space representation and performing inference in it. As discussed earlier in Section 3, the latent space accounts for the feature correlation information of the word embeddings. Moreover, the proposed factorization provides GeoMM the advantage to naturally extend to multilingual setting by learning to represent the word embeddings of multiple languages in a common latent space.

[9] Results for en-it and it-en have not been reported for CSLS-Fr-FV by Joulin et al. (2018).

Table 6: Pearson correlation coefficient for the SemEval 2017 cross-lingual word similarity task.

Method	**en-es**	**en-de**	**en-it**	**avg.**
NASARI	0.64	0.60	0.65	0.63
Luminoso_run2	0.75	0.76	0.77	0.76
Procrustes	0.72	0.72	0.71	0.72
MSF	0.73	0.74	0.73	0.73
GeoMM	0.73	0.74	0.74	0.74

In the following, we evaluate the effect of each component of the factorization, namely, the language specific rotation and the Mahalanobis similarity metric, in our algorithm:

- **No similarity metric is learned.** We enforce $\mathbf{B} = \mathbf{I}$, *i.e.,* the Mahalanobis similarity metric is not learned. Hence, we have $\mathbf{W} = \mathbf{U}_t \mathbf{U}_s^\top$, *i.e.,* $\mathbf{W}$ is also an orthogonal matrix. We observe that the results are abysmal, implying that learning an orthogonal transformation matrix does not seem to be sufficient in this case. It should be noted that this is a classification setting.
- **No language specific rotations are learned.** We enforce $\mathbf{U}_s = \mathbf{I}$ and $\mathbf{U}_t = \mathbf{I}$ in problem (4), *i.e.,* learn only the Mahalanobis metric $\mathbf{B}$ and $\mathbf{W} = \mathbf{B}$, where $\mathbf{B}$ is a symmetric positive definite matrix. We see a significant drop in performance, which suggests that the feature space of different languages needs to be *aligned* with each other via language specific orthogonal transformation.

To summarize, the classification setting is more suitable than the regression setting for solving the bilingual mapping problem in our framework. In addition, the similarity metric, the language specific rotations, and the inference in the latent space, together contribute in obtaining state-of-the-art performance. This shows that the proposed modeling choices are better than the alternatives compared in the ablation tests.

6.3 Cross-lingual Word Similarity

In addition to BLI, we evaluate our algorithm on the cross-lingual word similarity task using the SemEval 2017 dataset (Camacho-Collados et al., 2017). The results on the cross-lingual word similarity task are in Table 6. Our algorithm GeoMM performs better than Procrustes and MSF. It is also better than the Semval 2017 baseline method NASARI (Camacho-Collados et al., 2016) and is competitive with luminoso_run2, the best reported system on this dataset by Speer and Lowry-Duda (2017). It should be noted that NASARI and luminoso_run2 use additional knowledge sources like BabelNet and ConceptNet for building their embeddings.

6.4 Monolingual Word Analogy

A desirable property of cross-lingual embeddings is that the monolingual distances are preserved after mapping to avoid performance degradation in monolingual tasks. Following Artetxe et al. (2016), we test the monolingual quality of the mapped English embeddings (from Italian) on the monolingual analogy task of Mikolov et al. (2013a). We quote the

Table 7: Accuracy on the monolingual word analogy task.

Method	**Accuracy (%)**
Original English embeddings	76.66
Procrustes (Artetxe et al., 2016)	76.66
MSF	76.59
GeoMM	75.21

performance of the Procrustes solution reported by Artetxe et al. (2016). We also compare with the original monolingual embeddings as well as with MSF algorithm. Table 7 shows the results of the monolingual analogy task. We observe that there is no significant drop in the monolingual performance compared to 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 bilingual lexicon induction when a direct bilingual dictionary between the source and target language is available. In this section, we explore if the inherent multilingual nature of our formulation 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 GeoMM 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

To evaluate the multilingual embeddings, we consider the problem of BLI 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. As discussed in Section 3.3, the proposed framework can learn a mapping from L_{src} to L_{tgt} in this setting.

We adapt any supervised bilingual approach to the one-hop translation setting by considering their following variants:

- Composition (cmp): Using the given bilingual approach, we learn the individual $L_{\text{src}} \rightarrow L_{\text{pvt}}$ and $L_{\text{pvt}} \rightarrow L_{\text{tgt}}$ transformations as $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. Given a word embedding x from language L_{src}, the corresponding embedding in L_{tgt} is obtained by a composition of the individual transformations $\mathbf{W}_2\mathbf{W}_1x$. This is equivalent to compute the similarity of source and target language embeddings in the pivot language embedding space. Recently, Smith et al. (2017b) explored this technique with the Procrustes solution for $\mathbf{W}_1$ and $\mathbf{W}_2$. The algorithms under this variant include Procrustes (cmp), MSF (cmp), and GeoMM (cmp).
- Pipeline (pip): Using the given bilingual approach, we learn the individual $L_{\text{src}} \rightarrow L_{\text{pvt}}$ and $L_{\text{pvt}} \rightarrow L_{\text{tgt}}$ transformations as $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. Given a word embedding x from language L_{src}, we find the translation z in L_{pvt} perform inference on $\mathbf{W}_1x$. Let $z = \arg\max_i \text{CSLS}(\mathbf{W}_1x, z_i)$, where z_i represents the word embeddings from the

Table 8: Precision@1 for BLI for the indirect translation scenario.

Method	**it-en-de**	**de-en-es**	**fr-it-pt**	**it-de-es**	**es-pt-fr**	**avg.**
SL-unsup	**50.5**	42.5	74.1	86.4	84.6	67.6
Procrustes (cmp)	43.6	40.0	76.2	83.9	83.0	65.3
Procrustes (pip)	33.6	34.6	72.9	64.9	80.1	57.2
Procrustes (int)	45.7	42.2	77.5	86.7	84.9	67.4
MSF (cmp)	36.2	38.3	78.1	84.1	83.7	64.1
MSF (pip)	35.2	37.0	76.2	67.3	83.0	59.7
MSF (int)	46.4	42.1	80.3	**88.7**	86.9	68.9
GeoMM (cmp)	43.4	38.8	78.7	86.5	84.9	66.5
GeoMM (pip)	36.4	37.9	76.7	66.7	81.9	59.9
GeoMM (int)	48.5	43.9	80.4	**88.7**	86.3	69.6
GeoMM$_{\text{multi}}$	49.4	43.5	81.1	88.3	86.9	69.8
GeoMM$_{\text{multi+int}}$	50.3	**44.7**	**81.2**	88.7	**86.9**	**70.4**

vocabulary of language L_{pvt} and CSLS represents the cross-domain similarity local scaling score proposed by Conneau et al. (2018). Then, the corresponding embedding of x in L_{tgt} is $\mathbf{W}_2 z$. The algorithms under this variant include Procrustes (pip), MSF (pip), and GeoMM (pip).

- Intersection (int): We create a dictionary between L_{src} and L_{tgt} from the given two dictionaries. However, this is possible only when both dictionaries have some common words from L_{pvt}. We term this dictionary between L_{src} and L_{tgt} as the *intersection dictionary*. The algorithms under this variant include Procrustes (int), MSF (int), and GeoMM (int).

In addition to the above three variants of GeoMM, our framework allows the flexibility to jointly learn the common latent space of multiple languages, given bilingual dictionaries of multiple language pairs (Section 6). In particular, we also consider the following multilingual variants of our GeoMM approach:

- GeoMM$_{\text{multi}}$: only L_{src}-L_{pvt} and L_{pvt}-L_{tgt} dictionaries are available.
- GeoMM$_{\text{multi+int}}$: the L_{src}-L_{pvt} and L_{pvt}-L_{tgt} dictionaries as well as the intersection dictionary are available.

7.2 Experimental Settings

We experiment with the following one-hop translation cases: (a) it-en-de, (b) de-en-es, (c) fr-it-pt, (d) it-de-es, and (e) es-pt-fr (read the triplets as L_{src}-L_{pvt}-L_{tgt}). The training dictionaries as well as all the word embeddings for (a) and (b) are from VecMap dataset, while their test dictionaries (used for evaluating the translation quality) are from the MUSE dataset. This is because of the absence of dictionaries for it-de and de-es in the VecMap dataset. Cases (c)-(e) use datasets and embeddings from the MUSE dataset.

7.3 Results and Analysis

Table 8 shows the results of the one-hop translation experiments. We observe that $\text{GeoMM}_{\text{multi}}$ outperforms pivoting methods (cmp and pip) built on top of MSF and Procrustes for all language pairs. Pivoting may lead to cascading of errors in the pipeline, whereas learning a common embedding space jointly will not be prone to this disadvantage. This is reaffirmed by our observation that $\text{GeoMM}_{\text{multi}}$ performs significantly better than GeoMM (cmp) and GeoMM (pip) as well.

Learning a bilingual mapping from a small intersection dictionary is better the pivoting methods. However, $\text{GeoMM}_{\text{multi}}$ is also better than the use of an intersection dictionary to learn the bilingual mappings. The intersection dictionary method has to discard a lot of the available mappings, but $\text{GeoMM}_{\text{multi}}$ is able to leverage all the available data.

When the intersection dictionary is also used during training in addition to the original bilingual dictionaries ($\text{GeoMM}_{\text{multi+int}}$), we see a good improvement in the performance. Direct translation evidence reinforces the indirect translation evidence provided in the case of $\text{GeoMM}_{\text{multi}}$.

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 though it used no bilingual dictionaries. On the other hand, $\text{GeoMM}_{\text{multi}}$ is better than the unsupervised methods in most cases. It should be noted that the unsupervised methods use much larger vocabulary than $\text{GeoMM}_{\text{multi}}$ during the training stage.

8. Conclusion

In this work, we develop a framework for learning bilingual mapping for word embeddings by aligning the embeddings of the source and target language 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 learns the language specific orthogonal transformations and the Mahalanobis metric. We formulate the problem in the Riemannian optimization framework, which models the involved orthogonal and symmetric positive definite constraints efficiently. The evaluations on the bilingual lexicon induction task and cross-lingual word similarity task show that our approach outperforms existing approaches on multiple languages and datasets. Our analysis shows that the combination of learning a metric induced common latent space and modeling the problem in classification setting is beneficial for the bilingual lexicon induction task.

Our framework easily generalizes to the multilingual setting as it can represent multiple languages in a common latent space. We demonstrate this capability by evaluating our multilingual approach for bilingual lexicon induction in a one-hop translation setting. We show that our multilingual embeddings perform better than indirect translation using a pivot language, which has bilingual dictionaries with source and target languages.

References

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

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

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

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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, 2018a.

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

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

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

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

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

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, 2017.

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

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

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

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

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

Manaal Faruqui and Chris Dyer. 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, 2014.

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

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

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 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*, 2018.

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

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

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

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

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

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, and Edouard Grave. Improving supervised bilingual mapping of word embeddings. Technical report, arXiv preprint arXiv:1804.07745, 2018.

Michel Journée, Francis Bach, Pierre-Antoine Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

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

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 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, 2015.

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

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

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

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

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

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

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

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

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Aligning the fastText vectors of 78 languages, 2017b. URL `https://github.com/Babylonpartners/fastText_multilingual`.

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

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

Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, 81(1):21–35, Oct 2010.

Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2764–2770, 2011.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 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, 2015.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1959–1970, 2017a.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Earth movers distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, 2017b.

Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 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, 2015.