Geometry-aware Domain Adaptation for Unsupervised Alignment of Word Embeddings

Pratik Jawanpuria, Mayank Meghwanshi, Bamdev Mishra
Microsoft, India
{pratik.jawanpuria,mamegh,bamdevm}@microsoft.com

Abstract

We propose a novel manifold based geometric approach for learning unsupervised alignment of word embeddings between the source and the target languages. Our approach formulates the alignment learning problem as a domain adaptation problem over the manifold of doubly stochastic matrices. This viewpoint arises from the aim to align the second order information of the two language spaces. The rich geometry of the doubly stochastic manifold allows to employ efficient Riemannian conjugate gradient algorithm for the proposed formulation. Empirically, the proposed approach outperforms state-of-the-art optimal transport based approach on the bilingual lexicon induction task across several language pairs. The performance improvement is more significant for distant language pairs.

1 Introduction

Learning bilingual word embeddings is an important problem in natural language processing (Mikolov et al., 2013; Faruqui and Dyer, 2014; Artetxe et al., 2016; Conneau et al., 2018), with usage in cross-lingual information retrieval (Vulić and Moens, 2015), text classification (Wan et al., 2011; Klementiev et al., 2012), machine translation (Artetxe et al., 2018c) etc. Given a source-target language pair, the aim is to represent the words in both languages in a common embedding space. This is usually achieved by learning a linear function that maps word embeddings of one language to the embedding space of the other language (Mikolov et al., 2013).

Several works have focused on learning such bilingual mapping in supervised setting, using a bilingual dictionary during the training phase (Artetxe et al., 2018a; Joulin et al., 2018; Jawanpuria et al., 2019). Recently, unsupervised bilingual word embeddings have also been explored (Zhang et al., 2017a,b; Conneau et al., 2018; Artetxe et al., 2018b; Hoshen and Wolf, 2018; Grave et al., 2019; Alvarez-Melis and Jaakkola, 2018; Zhou et al., 2019; Jawanpuria et al., 2020).

Learning unsupervised cross-lingual mapping may be viewed as an instance of the more general unsupervised domain adaptation problem (Ben-David et al., 2007; Gopalan et al., 2011; Sun et al., 2016; Mahadevan et al., 2018). The latter fundamentally aims at aligning the input feature (embeddings) distributions of the source and target domains (languages). In this paper, we take this point of view and learn cross-lingual word alignment by finding alignment between the second order statistics of the source and the target language embedding space.

We formulate a novel optimization problem on the set of doubly stochastic matrices. The objective function consists of matching covariances of words from source to target languages in a least-squares sense. For optimization, we exploit the fact that the set of doubly stochastic matrices has rich geometry and forms a Riemannian manifold (Douik and Hassibi, 2019). The Riemannian optimization framework (Absil et al., 2008; Edelman et al., 1998; Smith, 1994) allows to propose a computationally efficient conjugate gradient algorithm (Douik and Hassibi, 2019). Experiments show the efficacy of the proposed approach on the bilingual lexicon induction benchmark, especially on the language pairs involving distant languages.

2 Motivation and Related Work

We introduce the bilingual word alignment setup followed by a discussion on domain adaptation approaches.

Bilingual alignment. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Z} \in \mathbb{R}^{n \times d}$ be d-dimensional word embeddings of n words of the source and the target languages, re-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3052–3058
July 5 - 10, 2020.

spectively. The aim is to learn a linear operator $\mathbf{W} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that best approximates source embeddings in the target language space.

In the supervised setup, a list of source words and their translations in the target language is provided. This is represented by an *alignment* matrix $\mathbf{Y}$ of size $n \times n$, where $\mathbf{Y}_{ij} = 1$ if j-th word in the target language is a translation of the i-th word in the source language and $\mathbf{Y}_{ij} = 0$ otherwise. A standard way to learn orthogonal $\mathbf{W}$ is by solving the *orthogonal Procrustes* problem (Artetxe et al., 2016; Smith et al., 2017), i.e.,

$$\begin{aligned} \min_{\mathbf{W} \in \mathbb{R}^{d \times d}} \quad & \|\mathbf{XW} - \mathbf{YZ}\|_{\text{Fro}}^2 \\ \text{subject to} \quad & \mathbf{W}^\top \mathbf{W} = \mathbf{I}, \end{aligned} \tag{1}$$

where $\|\cdot\|_{\text{Fro}}$ is the Frobenius norm and $\mathbf{I}$ is the identity matrix. Problem (1) has the closed-form solution $\mathbf{W}^\star = \mathbf{UV}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are the respective left and right orthogonal factors of the singular value decomposition of $\mathbf{X}^\top \mathbf{YZ}$ (Schönemann, 1966).

In the unsupervised setting, $\mathbf{Y}$ is additionally unknown apart from $\mathbf{W}$. Most unsupervised works (Zhang et al., 2017b; Artetxe et al., 2018b; Grave et al., 2019; Conneau et al., 2018) tackle this challenge by learning $\mathbf{Y}$ and $\mathbf{W}$ jointly. However, their performance rely on finding a good initialization candidate for the alignment matrix $\mathbf{Y}$ (Zhang et al., 2017b; Grave et al., 2019; Alaux et al., 2019; Jawanpuria et al., 2020).

Performing optimization over the set of binary matrices, $\mathbf{Y} \in \{0, 1\}^{n \times n}$, to learn the bilingual alignment matrix is computationally hard. Hence, some works (Zhang et al., 2017b; Xu et al., 2018) view the source and the target word embedding spaces as two distributions and learn $\mathbf{Y}$ as the transformation that makes the two distributions close. This viewpoint is based on the theory of *optimal transport* (Villani, 2009; Peyré and Cuturi, 2019). $\mathbf{Y}$ is, thus, modeled as a *doubly stochastic* matrix: the entries in $\mathbf{Y} \in [0, 1]$ and each row/column sums to 1. Permutation matrices are extreme points in the space of doubly stochastic matrices.

Alvarez-Melis and Jaakkola (2018) propose learning the doubly stochastic $\mathbf{Y}$ as a transport map between the *metric spaces* of the words in the source and the target languages. They optimize the Gromov-Wasserstein (GW) distance, which measures how distances between pairs of words are mapped across languages. For learning $\mathbf{Y}$, they propose to

$$\min_{\mathbf{Y} \in \mathbb{DS}^n} \quad -\text{Trace}(\mathbf{Y}^\top \mathbf{C_X} \mathbf{Y} \mathbf{C_Z}), \tag{2}$$

where $\mathbb{DS}^n := \{\mathbf{Y} \in \mathbb{R}^{n \times n} : \mathbf{Y} \geq 0, \mathbf{Y}^\top \mathbf{1} = \mathbf{1} \text{ and } \mathbf{Y1} = \mathbf{1}\}$ is the set of $n \times n$ doubly stochastic matrices, $\mathbf{Y} \geq 0$ implies entry-wise non-negativity, $\mathbf{1}$ is a column vector of ones, and $\mathbf{C_X} = \mathbf{XX}^\top$ and $\mathbf{C_Z} = \mathbf{ZZ}^\top$ are $n \times n$ word covariance matrices of source and target languages, respectively. An iterative scheme is proposed for solving (2), where each iteration involves solving an optimal transport problem with *entropic regularization* (Peyré et al., 2016; Peyré and Cuturi, 2019). The optimal transport problem is solved with the popular *Sinkhorn* algorithm (Cuturi, 2013). It should be noted that the GW approach (2) only learns $\mathbf{Y}$. The linear operator to map source language word embedding to the target language embedding space can then be learned by solving (1).

Domain adaptation. Domain adaption refers to transfer of information across domains and has been an independent research of interest in many fields including natural language processing (Daumé III, 2007; Borgwardt et al., 2006; Adel et al., 2017; Baktashmotlagh et al., 2013; Fukumizu et al., 2007; Wang et al., 2015; Prettenhofer and Stein, 2011; Wan et al., 2011; Sun et al., 2016; Mahadevan et al., 2018; Ruder, 2019).

One modeling of interest is by Sun et al. (2016), who motivate a linear transformation on the features in source and target domains. In (Sun et al., 2016), the linear map $\mathbf{A} \in \mathbb{R}^{d \times d}$ is solved by

$$\min_{\mathbf{A} \in \mathbb{R}^{d \times d}} \quad \left\|\mathbf{A}^\top \mathbf{D_X} \mathbf{A} - \mathbf{D_Z}\right\|_{\text{Fro}}^2, \tag{3}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are $d \times d$ are feature covariances of source and target domains (e.g., $\mathbf{D_X} = \mathbf{X}^\top \mathbf{X}$ and $\mathbf{D_Z} = \mathbf{Z}^\top \mathbf{Z}$), respectively. Interestingly, (3) has a closed-form solution and shows good performance on standard benchmark domain adaptation tasks (Sun et al., 2016).

3 Domain Adaptation Based Cross-lingual Alignment

The domain adaptation solution strategies of (Sun et al., 2016; Mahadevan et al., 2018) can be motivated directly for the cross-lingual alignment problem by dealing with word covariances instead of feature covariances. However, the cross-lingual word alignment problem additionally has a bi-directional symmetry: if $\mathbf{Y}$ aligns $\mathbf{X}$ to $\mathbf{Z}$, then

$\mathbf{Y}^\top$ aligns $\mathbf{Z}$ to $\mathbf{X}$. We exploit this to propose a bi-directional domain adaptation scheme based on (3). The key idea is to adapt the second order information of the source and the target languages into each other's domain. We formulate the above as follows:

$$\min_{\mathbf{Y}\in\mathbb{DS}^n} \quad \|\mathbf{Y}^\top\mathbf{C_X}\mathbf{Y} - \mathbf{C_Z}\|_{\text{Fro}}^2 + \|\mathbf{Y}\mathbf{C_Z}\mathbf{Y}^\top - \mathbf{C_X}\|_{\text{Fro}}^2, \tag{4}$$

The first term in the objective function $\|\mathbf{Y}^\top\mathbf{C_X}\mathbf{Y} - \mathbf{C_Z}\|_{\text{Fro}}^2$ adapts the domain of $\mathbf{X}$ (source) into $\mathbf{Z}$ (target). Equivalently, minimizing only the first term in the objective function of (4) leads to row indices in $\mathbf{Y}^\top\mathbf{X}$ aligning closely with the row indices of $\mathbf{Z}$. Similarly, minimizing only the second term $\|\mathbf{Y}\mathbf{C_Z}\mathbf{Y}^\top - \mathbf{C_X}\|_{\text{Fro}}^2$ adapts $\mathbf{Z}$ (now treated as the source domain) into $\mathbf{X}$ (now treated as the target domain), which means that the row indices $\mathbf{YZ}$ and $\mathbf{X}$ are closely aligned. Overall, minimizing both the terms of the objective function allows to learn the alignment matrix $\mathbf{Y}$ from $\mathbf{X}$ to $\mathbf{Z}$ and $\mathbf{Y}^\top$ from $\mathbf{Z}$ to $\mathbf{X}$ simultaneously. Empirically, we observe that bi-directionality acts as a self regularization, leading to optimization stability and better generalization ability.

The differences of the proposed formulation (4) with respect to the GW formulation (2) are two fold. First, the formulation (2) maximizes the inner product between $\mathbf{Y}^\top\mathbf{C_X}\mathbf{Y}$ and $\mathbf{C_Z}$. This inner product is sensitive to differences in the norms of $\mathbf{Y}^\top\mathbf{C_X}\mathbf{Y}$ and $\mathbf{C_Z}$. The proposed approach circumvents this issue since (4) explicitly penalizes entry-wise mismatch between $\mathbf{Y}^\top\mathbf{C_X}\mathbf{Y}$ and $\mathbf{C_Z}$. Second, the GW algorithm for (2) is sensitive to choices of the entropic regularization parameter (Alvarez-Melis and Jaakkola, 2018; Peyré and Cuturi, 2019). In our case, no such regularization is required.

Most recent works that solve optimal transport problem by optimizing over doubly stochastic matrices employ the Sinkhorn algorithm with entropic regularization (Cuturi, 2013; Peyré et al., 2016; Peyré and Cuturi, 2019). In contrast, we exploit the Riemannian manifold structure of the set of doubly stochastic matrices ($\mathbb{DS}^n$) recently studied in (Douik and Hassibi, 2019). $\mathbb{DS}^n$ is endowed with a smooth Fisher information metric (inner product) that makes the manifold smooth (Douik and Hassibi, 2019; Sun et al., 2015; Lebanon and Lafferty, 2004). In differential geometric terms, $\mathbb{DS}^n$ has the structure of a Riemannian submanifold. This makes computation of optimization-related ingredients, e.g., gradient and Hessian of a function, projection operators, and retraction operator, straightforward. Leveraging the versatile Riemannian optimization framework (Absil et al., 2008; Edelman et al., 1998; Smith, 1994), the constrained problem (4) is conceptually transformed to an *unconstrained* problem over the nonlinear manifold. Consequently, most unconstrained optimization algorithms generalize well to manifolds. We solve (4) using the Riemannian conjugate gradient algorithm (Absil et al., 2008; Douik and Hassibi, 2019).

There exist several manifold optimization toolboxes such as Manopt (Boumal et al., 2014), Pymanopt (Townsend et al., 2016), Manopt.jl (Bergmann, 2019), McTorch (Meghwanshi et al., 2018) or ROPTLIB (Huang et al., 2016), which have scalable off-the-shelf generic implementation of Riemannian algorithms. We use Manopt for our experiments, where we only need to provide the objective function (4) and its derivative with respect to $\mathbf{Y}$. The manifold optimization related ingredients are handled by Manopt internally. The computational cost per iteration of the algorithm is $O(n^2)$, which is similar to that of GW (Alvarez-Melis and Jaakkola, 2018).

We term our algorithm as **M**anifold **B**ased **A**lignment (MBA) algorithm. Our code is available at `https://pratikjawanpuria.com/publications/`.

4 Experiments

We compare the proposed algorithm MBA with state-of-the-art GW alignment algorithm (Alvarez-Melis and Jaakkola, 2018) for the bilingual induction (BLI) task. Both the algorithms use second order statistics (word covariance matrices) to learn the word alignment between two languages. In our experimental setup, we first learn the word alignment between the source and the target languages and then compute cross-lingual mapping by solving the Procrustes problem (1). For inference of nearest neighbors, we employ the cross-domain similarity local scaling (CSLS) similarity score (Conneau et al., 2018). We report Precision@1 (P@1) as in (Alvarez-Melis and Jaakkola, 2018; Artetxe et al., 2018b) for the BLI task.

We show results on the MUSE dataset (Conneau et al., 2018), which consists of fastText monolingual embeddings for different languages (Bojanowski et al., 2017) and dictionaries between several languages (but mostly with English). Follow-

Method	de-xx	en-xx	es-xx	fr-xx	it-xx	pt-xx	xx-de	xx-en	xx-es	xx-fr	xx-it	xx-pt	avg.
GW	62.6	77.4	**78.2**	**75.4**	**77.5**	77.2	62.6	75.9	**79.7**	**79.0**	**76.2**	74.9	74.7
MBA	**63.3**	**78.4**	**78.2**	75.3	77.0	**77.5**	**63.1**	**77.3**	79.4	78.7	**76.2**	**75.0**	**75.0**

Table 1: P@1 for BLI on six European languages: English, German, Spanish, French, Italian, and Portuguese. Here 'en-xx' refers to the average P@1 when English is the source language and others are target language. Similarly, 'xx-en' implies English as the target language and others as source language. Thus, 'avg.' shows P@1 averaged over all the thirty BLI results for each algorithm. The proposed algorithm MBA performs similar when the language pairs are closely related to each other.

Method	en-bg	en-cs	en-da	en-el	en-fi	en-hu	en-nl	en-pl	en-ru	
GW	22.8	42.1	54.4	21.5	37.7	43.7	72.9	49.1	36.1	
MBA	**38.1**	**46.8**	**56.1**	**40.0**	**40.4**	**46.1**	**73.8**	**50.4**	**37.5**	
Method	**bg-en**	**cs-en**	**da-en**	**el-en**	**fi-en**	**hu-en**	**nl-en**	**pl-en**	**ru-en**	**avg.**
GW	29.9	52.9	60.7	32.7	49.5	57.6	70.9	57.7	48.3	47.0
MBA	**50.0**	**57.7**	**62.3**	**54.4**	**54.4**	**61.0**	**71.0**	**60.5**	**54.1**	**53.0**

Table 2: P@1 for BLI on English and nine European languages: Bulgarian, Czech, Danish, Greek, Finnish, Hungarian, Dutch, Polish, and Russian. The 'avg.' shows P@1 averaged over all the eighteen BLI results. The proposed algorithm MBA outperforms GW when the bilingual mapping is learned between distant languages.

Method	en-ar	en-hi	en-tr	ar-en	hi-en	tr-en
GW	27.4	0.0	40.9	**41.0**	0.0	52.4
MBA	**27.9**	**25.1**	**42.0**	40.8	**28.9**	**54.6**

Table 3: P@1 for BLI on English and three non-European languages (Arabic, Hindi, and Turkish). MBA obtains significantly better results.

ing existing works (Artetxe et al., 2018b; Alvarez-Melis and Jaakkola, 2018; Alaux et al., 2019), the embeddings are normalized. The MUSE dataset provides predefined thirty test bilingual dictionaries between six European languages: English (en), German (de), Spanish (es), French (fr), Italian (it), and Portuguese (pt) on which we evaluate the methods. Additionally, we compute performance on the test dictionaries between English and twelve other languages: Arabic (ar), Bulgarian (bg), Czech (cs), Danish (da), Dutch (nl), Finnish (fi), Greek (el), Hindi (hi), Hungarian (hu), Polish (po), Russian (ru), and Turkish (tr). Following Alvarez-Melis and Jaakkola (2018), we consider top $n = 20\,000$ most frequent words in the vocabulary set for all the languages during the training stage. The inference is performed on the the full vocabulary set.

For GW, we use the original codes shared by Alvarez-Melis and Jaakkola (2018) and follow their recommendations on tuning the entropic regularization parameter and scaling of covariance matrices $\mathbf{C_X}$ and $\mathbf{C_Z}$. As a practical implementation of MBA, we incrementally increase n starting from 1000 to 20 000 every fixed-number of iterations.

We begin by discussing the results on six close-by European languages in Table 1. We observe that both MBA and GW perform similarly when the languages are related. Hence, in the second set of experiments, we consider other European languages that are distant to English. We observe from Table 2 that MBA outperforms GW, by an average BLI score of 6 points, in this challenging setting. Table 3 reports results on language pairs involving English and three non-European languages. We again observe that the proposed algorithm MBA performs significantly better than GW. Overall, the experiments show the benefit of a geometric optimization framework.

5 Conclusion

Aligning the metric spaces of languages has a wide usage in cross-lingual applications. A popular approach in literature is the Gromov-Wasserstein (GW) alignment approach (Mémoli, 2011; Peyré et al., 2016; Alvarez-Melis and Jaakkola, 2018), which constructs a transport map by viewing the two embedding spaces as distributions. In contrast, we have viewed unsupervised bilingual word alignment as an instance of the more general unsupervised domain adaptation problem. In particular, our formulation allows search over the space of doubly stochastic matrices and induces bi-directional mapping between the source and target words. Both are motivated solely from the language perspective.

The Riemannian framework allows to exploit the geometry of the doubly stochastic manifold. Empirically, we observe that the proposed algorithm MBA outperforms the GW algorithm for learning bilingual mapping (Alvarez-Melis and Jaakkola, 2018), demonstrating the benefit of geometric optimization modeling.

References

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

Tameem Adel, Han Zhao, and Alexander Wong. 2017. Unsupervised domain adaptation with a relaxed covariate shift assumption. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyper-alignment for multilingual word embeddings. In *Proceedings of the International Conference on Learning Representations*. URL: https://github.com/facebookresearch/fastText/tree/master/alignment.

David Alvarez-Melis and Tommi S. Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. URL: https://github.com/dmelis/otalign.

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. 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. URL: https://github.com/artetxem/vecmap.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018c. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642.

Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. 2013. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144.

Ronny Bergmann. 2019. Optimisation on Manifolds in Julia. https://github.com/kellertuer/Manopt.jl.

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

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57.

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.

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*. URL: https://github.com/facebookresearch/MUSE.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Ahmed Douik and Babak Hassibi. 2019. Manifold optimization over the set of doubly stochastic matrices: A second-order geometry. *IEEE Transactions on Signal Processing*, 67(22):5761–5774.

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.

Kenji Fukumizu, Francis R Bach, and Arthur Gretton. 2007. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(Feb):361–383.

R. Gopalan, Ruonan Li, and R. Chellappa. 2011. Domain adaptation for object recognition: An unsupervised approach. In *International Conference on Computer Vision*.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *International Conference on Artificial Intelligence and Statistics*.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. Technical report, arXiv preprint arXiv:1801.06126v3.

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.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: A geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120.

Pratik Jawanpuria, Mayank Meghwanshi, and Bamdev Mishra. 2020. A simple approach to learning unsupervised multilingual embeddings. Technical report, arXiv preprint arXiv:2004.05991.

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

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.

Guy Lebanon and John Lafferty. 2004. Hyperplane margin classifiers on the multinomial manifold. In *Proceedings of the International Conference on Machine learning*, page 66.

Sridhar Mahadevan, Bamdev Mishra, and Shalini Ghosh. 2018. A unified framework for domain adaptation using metric learning on manifolds. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 843–860.

Mayank Meghwanshi, Pratik Jawanpuria, Anoop Kunchukuttan, Hiroyuki Kasai, and Bamdev Mishra. 2018. Mctorch, a manifold optimization library for deep learning. Technical report, arXiv preprint arXiv:1810.01811.

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

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

Gabriel Peyré and Marco Cuturi. 2019. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–602.

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

Peter Prettenhofer and Benno Stein. 2011. Cross-lingual adaptation using structural correspondence learning. *ACM Transactions on Intelligent Systems and Technology*, 3(1):13.

Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Ph.D. thesis, National University of Ireland, Ireland.

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

S. T. Smith. 1994. Optimization techniques on Riemannian manifold. In A. Bloch, editor, *Hamiltonian and Gradient Flows, Algorithms and Control*, volume 3, pages 113–136. American Mathematical Society, Providence, RI.

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

Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yanfeng Sun, Junbin Gao, Xia Hong, Bamdev Mishra, and Baocai Yin. 2015. Heterogeneous tensor decomposition for clustering via manifold optimization. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):476–489.

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.

Cédric Villani. 2009. *Optimal Transport: Old and New*. Springer-Verlag.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372.

Chang Wan, Rong Pan, and Jiefei Li. 2011. Bi-weighting domain adaptation for cross-language text classification. In *International Joint Conference on Artificial Intelligence*.

Weiran Wang, Raman Arora, Karen Livescu, and Nathan Srebro. 2015. Stochastic optimization for deep cca via nonlinear orthogonal iterations. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 688–695.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density matching for bilingual word embedding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.