Leveraging Smooth Manifolds for Lexical Semantic Change Detection across Corpora

Anmol Goel
GGSIPU, Delhi
agoel00@gmail.com

Ponnurangam Kumaraguru
IIIT-Delhi
pk@iiitd.ac.in

Abstract

Comparing two bodies of text and detecting words with significant lexical semantic shift between them is an important part of digital humanities. Traditional approaches have relied on aligning the different embeddings in the Euclidean space using the Orthogonal Procrustes problem. This study presents a geometric framework that leverages optimization on smooth Riemannian manifolds for obtaining corpus-specific orthogonal rotations and a corpus-independent scaling to project the different vector spaces into a shared latent space. This enables us to capture any affine relationship between the embedding spaces while utilising the rich geometry of smooth manifolds. By converting a constrained optimization problem to an unconstrained optimization problem directly on the manifold, we are able to exploit the structural bias of smooth manifolds to capture latent relationships between the embedding spaces.

1 Introduction

Investigating corpora from different sources by identifying words with semantic variations across them is integral to computational social science. Lexical Semantic Change (LSC) detection has recently seen a rise in interest due to the increasing effectiveness of word embeddings to encode various aspects of language. LSC Detection has been applied in two broad paradigms - diachronic (corpora from different time periods) and synchronic (corpora from different domains like populations, geographies, etc.) tasks. This is important in various use cases such as historical language processing, analyzing cross domain ambiguity and studying how language changes over time and across cultures.

The most common approach to this task involves training word embeddings on each corpus C_1 and C_2 from sources s_1 and s_2 respectively and then aligning the different embedding spaces using an alignment algorithm such that the distance between the embedding of a word w from each vector space can be used as a measure of its semantic change. A popular algorithm for this alignment is the Orthogonal Procrustes (OP) method which was introduced by [3] and has been shown as the best performing among several baselines [8]. The OP method solves the following constrained optimization problem

$$R = \underset{Q^TQ=I}{\operatorname{argmin}} ||QX - Y||_2$$

where X and Y are the word embeddings matrices trained on different corpora.

In this study, we propose a geometric framework for aligning the vector spaces which leverages the rich geometry of smooth Riemannian manifolds. By using smooth manifolds we are able to convert the problem from a constrained optimization in Euclidean space to an unconstrained optimization on the manifold itself. Traditional optimization methods like gradient descent can work on smooth manifolds through retractions. With the help of corpus-specific orthogonal transformations we align the different vector spaces in a common latent space. Additionally, we learn a Mahalanobis metric to scale different features within the latent space.

NeurIPS 2020 workshop on Differential Geometry meets Deep Learning

2 Preliminaries

2.1 Manifold

A structure that can be locally approximated by the Euclidean space $\mathcal{R}^n$ is known as the manifold $\mathcal{M}$ of dimension n. Manifolds are a generalization of the concept of surfaces. An example of a manifold in $\mathcal{R}^3$ which can be locally approximated by $\mathcal{R}^2$ is the sphere $\mathcal{S}^2$.

2.2 Tangent Space

For any point $x \in \mathcal{M}$, the tangent space $T_x\mathcal{M}$ of $\mathcal{M}$ at x is the vector space approximating $\mathcal{M}$ around x. A vector $v \in T_x\mathcal{M}$ is called the tangent vector of $\mathcal{M}$.

2.3 Riemannian Metric

A Riemannian metric g on $\mathcal{M}$ defines an inner product

$$g_x : T_x\mathcal{M} \times T_x\mathcal{M} \rightarrow \ R$$

on $T_x\mathcal{M}$ for each $x \in \mathcal{M}$. This is the generalization of the concept of measuring lengths in Euclidean geometry.

The simplest example of a metric is the Euclidean metric g defined in $\mathcal{R}^n$ as:

$$g_x(v, w) = \sum_{i=1}^{n} v^i w^i = v \cdot w$$

The Stiefel manifold is equipped with the Frobenius inner product as its Riemannian metric.

2.4 Riemannian Manifold

Using the concepts discussed so far, we can arrive at the formal definition of a smooth Riemannian manifold.

A smooth manifold equipped with a Riemannian metric is called a Riemannian manifold.

2.5 Manifold Optimization

We are concerned with optimization problems of the form

$$\underset{x \in \mathcal{M}}{\operatorname{argmin}} f(x)$$

where $f : \mathcal{M} \rightarrow \ R$, i.e., f is a real valued function on $\mathcal{M}$ and the search space $\mathcal{M}$ is smooth (differentiable manifold).

This formulation allows us to convert an optimization problem in the Euclidean space with constraints (like orthogonality of matrix) to an optimization where we directly work on the Riemannian manifolds.

We can perform gradient descent over the parameters lying on such manifolds using exponential maps which map a point on $T_x\mathcal{M}$ back to $\mathcal{M}$. For a more mathematical rigorous approach to these definitions, we point the reader to [9].

3 Proposed Approach

Our proposed approach closely follows previous work [4] from the cross lingual embeddings literature. We introduce this framework in the context of LSC Detection.

Consider two embeddings $A \in \mathcal{R}^{d \times n}$ and $B \in \mathcal{R}^{d \times n}$, trained on corpora C_1 and C_2 from different sources s_1 and s_2, where d is the dimensionality of embeddings and n is the number of words in the shared vocabulary, i.e., $n = |V|$. Our proposed approach learns corpus-specific orthogonal transformations $P \in \mathcal{O}^d$ and $Q \in \mathcal{O}^d$ to project A and B in a shared latent space. Here, $\mathcal{O}^d$ is the smooth manifold of $d \times d$ orthogonal matrices which is also known as the *Stiefel manifold*. Additionally, we induce the latent space with a Mahalanobis metric M to generalize the notion of cosine similarity. Mahalanobis metric captures the feature correlation information from the embedding matrices unlike cosine similarity. Here, M lies on the manifold of $d \times d$ symmetric positive definite matrices, i.e., $M \succ 0$.

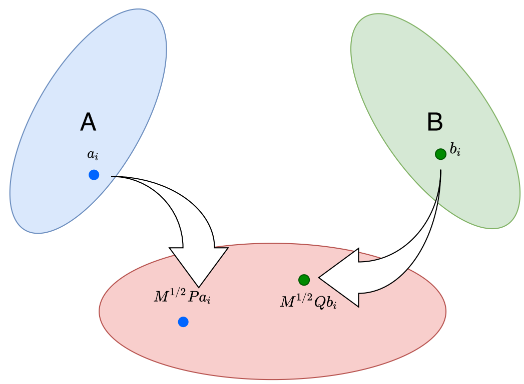

Figure 1: Illustrative example of the proposed transformations M, P and Q to transform the original embeddings in a common latent space

The objective function is supervised by aligning each word to itself, based on the assumption that most words across the two corpora remain stable or unchanged. Let Y be an identity matrix of size $n = |V|$ to denote the mapping between words from the two spaces, i.e., $Y_{ij} = 1$ for $i = j$ and 0 otherwise. We propose to solve the following optimization problem over the squared loss function:

$$\underset{P,Q \in \mathcal{O}^d; B \succ 0}{\text{argmin}} \quad ||A^T P^T M Q B - Y||^2 + \lambda ||M||^2$$

here λ is the regularization parameter and $||.||$ is the Frobenius norm.

In this induced latent space, similarity between $a_i \in A$ and $b_i \in B$ can be computed by the dot product $(Pa_i)^T M (Qb_i)$. As per [4], this is equivalent to the cosine similarity between $M^{\frac{1}{2}} Pa_i$ and $M^{\frac{1}{2}} Qb_i$. We use the popular Conjugate Gradient algorithm to solve the proposed optimization problem over the cartesian product of Stiefel and Positive Definite manifold which has a smooth Riemannian manifold structure. We utilise the widely studied square loss function and l_2-norm regularization in our proposed function.

Note that before transforming the embeddings, we take the intersection of their vocabulary so that only the words shared across corpora are considered. Figure 1 illustrates the learned latent space with our proposed approach from the original vector spaces trained on different corpora.

After aligning the embeddings using the proposed approach, for any given word $w_i \in |V|$, and its embeddings x_i and z_i from the original embedding spaces, we can measure its level of semantic change by computing the cosine distance between $M^{\frac{1}{2}} Px_i$ and $M^{\frac{1}{2}} Qz_i$, i.e.,

$$LCS_{w_i} = 1 - \frac{(M^{\frac{1}{2}} Px_i) \cdot (M^{\frac{1}{2}} Qz_i)}{||M^{\frac{1}{2}} Px_i||\,||M^{\frac{1}{2}} Qz_i||}$$

4 Implementation Details

We train word embeddings using the popular SGNS (word2vec) [6] approach on the DURel [8] corpus with dimension size 300, context window size 5 and all hyperparameters same as reported in the baseline paper [2] to keep the results consistent. We use 100 as the regularization parameter which has been shown to give good performance in previous studies [4]. Additionally, we L_2 normalize all the embeddings before optimization.

Recently, there has been increased support for Riemannian geometry thanks to publicly available geometric toolboxes like Geoopt [5], geomstats [7], manopt [1], etc. To optimise the proposed loss function on manifolds we use Pymanopt [10] library where we only need to provide the objective function and Riemannian optimization is handled automatically. In particular, we use the PositiveDefinite and Stiefel manifold implementations provided in the library to model the transformations and scaling operation. We utilise their implementation of the Riemannian Conjugate Gradient algorithm.

5 Quantitative Evaluation

We show the effectiveness of our approach on the gold standard diachronic corpora for LSC Detection evaluation: DURel introduced by [8]. DURel consists of 22 German words sampled from the DTA corpus with varying degrees of LSC manually annotated by human annotators by assigning each word in the list a degree of semantic change between 1 to 4. For training the word embeddings, we rely on the two splits of text corpora as used in previous studies: DTA documents from 1750-1799 and 1850-1899 in our experiment.

We compare our results with the state of the art approach of the stable word neighbors method studied by [2]. Based on the evaluation metric used in previous studies, we too compare the Spearman correlation between the predicted semantic change and human annotated degree of semantic change in the DURel dataset. The results are reported in Table 1. The superior performance of the proposed

Method	**Measure**	**DURel**
[2]	spearman	0.59
Proposed	spearman	**0.77**

Table 1: Results on diachronic corpus DURel

approach indicates that posing the alignment of vector spaces as a classification problem and learning corpus-specific transformations unlike previous approaches may result in better performance. Similar observations have been made previously by [4] in the cross lingual setting. The problem of LSC Detection may also show improvement in an unsupervised regime to bypass the self-contradicting objective problem as mentioned by [2]. We leave this for future work.

6 Conclusion and Future Work

In this study, we have introduced a geometric framework popularised in the cross lingual setting by [4], in the context of LSC Detection. This enables us to capture the latent relationships between the two embedding spaces by utilising the rich geometry of matrix manifolds. We hope to extend our approach by including contextual word embedding models in the rich geometry of smooth manifolds. Another possible future direction of our approach is to use the semantic change aware word embeddings as input features in downstream tasks like named entity recognition, part-of-speech tagging, etc as this could potentially lead to improvements in performance when applied to historical text documents.

References

[1] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42):1455–1459, 2014.

[2] H. Gonen, G. Jawahar, D. Seddah, and Y. Goldberg. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, 2020.

[3] W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Aug. 2016.

[4] P. Jawanpuria, A. Balgovind, A. Kunchukuttan, and B. Mishra. Learning multilingual word embeddings in latent metric space: A geometric approach. *Transactions of the Association for Computational Linguistics*, pages 107–120, 2019.

[5] M. Kochurov, R. Karimov, and S. Kozlukov. Geoopt: Riemannian optimization in pytorch, 2020.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[7] N. Miolane, A. L. Brigant, J. Mathe, B. Hou, N. Guigui, Y. Thanwerdas, S. Heyder, O. Peltre, N. Koep, H. Zaatiti, H. Hajri, Y. Cabanes, T. Gerald, P. Chauchat, C. Shewmake, B. Kainz, C. Donnat, S. Holmes, and X. Pennec. Geomstats: A python package for riemannian geometry in machine learning, 2020.

[8] D. Schlechtweg, A. Hätty, M. del Tredici, and S. Schulte im Walde. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, 2019.

[9] M. Spivak. *A comprehensive introduction to differential geometry*, volume 4. 1979.

[10] J. Townsend, N. Koep, and S. Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016.