Detecting Lexical Semantic Change across Corpora with Smooth Manifolds (Student Abstract)

Anmol Goel [1], **Ponnurangam Kumaraguru** [2]
[1]Guru Gobind Singh Indraprastha University, Delhi
[2]IIIT-Delhi
agoel00@gmail.com, 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 using the Orthogonal Procrustes problem in the Euclidean space. This study presents a geometric framework that leverages smooth Riemannian manifolds for corpus-specific orthogonal rotations and a corpus-independent scaling metric 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.

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 (Hamilton, Leskovec, and Jurafsky 2016) which has been shown as the best performing among several baselines (Schlechtweg et al. 2019). 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 embedding 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.

Proposed Approach

Our proposed approach closely follows previous work (Jawanpuria et al. 2019) 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., $M \succ 0$. 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:

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

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 (Jawanpuria et al. 2019), 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.

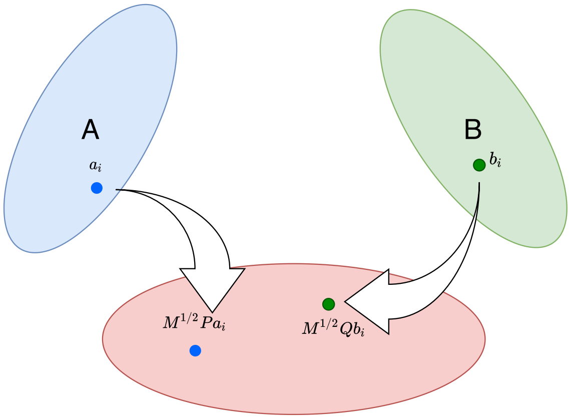

Figure 1: Illustrative example of the proposed transformations M, P and Q to transform the original embeddings in a common latent space

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.,

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

Quantitative Evaluation

We show the effectiveness of our approach by training word2vec embeddings on the gold standard diachronic corpora for LSCD evaluation: DURel (Schlechtweg et al. 2019). 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 use dimension size 300, context size 5 and keep all other hyperparameters same as (Gonen et al. 2020).

We compare our results with the state of the art approach of the stable word neighbors method studied by (Gonen et al. 2020). 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 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 (Jawanpuria et al. 2019) 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 (Gonen et al. 2020). We leave this for future work.

Method	**Measure**	**DURel**
(Gonen et al. 2020)	spearman	0.59
Proposed	spearman	**0.77**

Table 1: Results on diachronic corpus DURel

Conclusion and Future Work

In this study, we have introduced a geometric framework popularised in the cross lingual setting by (Jawanpuria et al. 2019), in the context of LSC Detection. 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. We hope to extend our approach by including contextual word embedding models in the rich geometry of smooth manifolds.

References

Gonen, H.; Jawahar, G.; Seddah, D.; and Goldberg, Y. 2020. 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*, 538–555.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016. 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)*, 1489–1501.

Jawanpuria, P.; Balgovind, A.; Kunchukuttan, A.; and Mishra, B. 2019. Learning Multilingual Word Embeddings in Latent Metric Space: A Geometric Approach. *Transactions of the Association for Computational Linguistics* 107–120.

Schlechtweg, D.; Hätty, A.; del Tredici, M.; and Schulte im Walde, S. 2019. 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*, 732–746.