On Approximately Orthogonal Matrices in Bilingual Word Embedding Mapping

鄧 一凡 * 呂 照陽 † 鶴岡 慶雅 *
東京大学 情報理工学系研究科
{dengyifan, tsuruoka}@logos.t.u-tokyo.ac.jp
lyuzhaoyang@link.cuhk.edu.hk

1 Introduction

Bilingual word embedding alignment is widely applied in bilingual natural language processing (NLP) tasks, such as bilingual lexicon induction (BLI) [3] and unsupervised sentence translation [5]. The reason why bilingual word embedding alignment can be successful lies in the phenomenon that word embedding spaces share similar geometric arrangements discovered by Mikolov et al. [5]. Xing et al. [10] propose to normalize the word embeddings and apply an orthogonal matrice to map the source embeddings into a normalized target embedding space, which is widely adopted by later studies.

However, Patra et al. [6] argue that strictly orthogonal mapping is not optimal for bilingual word embedding alignment, since Søgaard et al. [9] have shown that the word embedding spaces are not isomorphic. They apply a relaxed orthogonal constraint on the mapping matrix to obtain an approximately orthogonal mapping matrix. The relaxed constraint is utilized along with other constraints in their study. In this paper, we investigate the effect of the approximate orthogonality of the mapping matrix and propose approximate orthogonality refinement based on our investigation.

2 Related Work

2.1 Bilingual word embedding alignment

In the supervised setting of bilingual word embedding alignment, there is an initial bilingual word lexicon. Mikolov et al. [5] propose to optimize a tansformation matrix W:

$$W^* = \underset{W}{\arg\min} \sum_{i=1}^{n} \|Wx_i - z_i\|^2, \tag{1}$$

where $\|\cdot\|$ denotes the Frobenius norm, x_i and z_i are respective word embeddings of a word pair in the initial lexicon, and n is the size of the lexicon.

*The University of Tokyo
†The Chinese University of Hong Kong

Xing et al. [10] propose to orthogonalize the mapping matrix and then apply the cosine similarity between the mapped embedding and the target embedding as the target, resulting in the following optimization problem:

$$W^* = \underset{W}{\arg\max} \sum_{i=1}^{n} (Wx_i)^T z_i. \tag{2}$$

Note that by normalizing the embeddings and orthogonalize the mapping matrix, the cosine similarity can be calculated with the inner product.

Besides Euclidean distance and cosine similarity, there are also other distant measures used as training targets, such as earth mover's distance [12] and relaxed cross-domain similarity local scaling (RCSLS) [4].

Bilingual word embedding alignment is usually evaluated with BLI. For BLI, a lexicon is generated by searching the nearest neighbor of the mapped embedding in the target language embedding space. Then, the generated lexicon is compared with a gold lexicon – the more similar the better. The nearest neighbor is usually extracted by cosine distance. However, due to the hubness problem [7] in high dimension spaces that some points are the nearest neighbor of many points, Conneau et al. propose to search the nearest neighbor by cross-domain similarity local scaling (CSLS) [3], which usually results in better BLI performance than searching by cosine distance.

2.2 Orthogonality constraints

Xing et al. [10] propose to normalize the word embeddings and apply an orthogonal matrix to map the source embeddings into the normalized target embedding space, which is shown to have better performance over unconstrainted linear mapping matrices. Artetxe et al. [1] report that by orthogonal mapping matrices, monolingual features of the mapped embedding would be best maintained. Smith et al. [8] proved that linear mapping is self-consistent only when the matrix is orthogonal. Orthogonal matrices have other desirable properties, e.g. their transpose can perform back transform [12, 11], and orthogonality makes the training procedure stable [3]. With the aforementioned properties, orthogonal matrices have been popular in former studies.

Former studies have been using different ways to constrain the mapping matrix to be orthogonal. Some approaches create strictly orthogonal matrices and others create approximately orthogonal matrices.

Strong orthogonal constraint

The main approach to force the mapping matrix to be orthogonal is to solve the orthogonal Procrustes problem [1]. The problem is formulated as

$$R = \arg\min_{\Omega} \|\Omega A - B\| \tag{3}$$

$$\text{s.t.} \quad \Omega^{\mathrm{T}}\Omega = \mathrm{I}. \tag{4}$$

The orthogonal Procrustes problem is the same as the supervised setting of bilingual word embedding mapping if we see A as the source language embeddings in the dictionary, B as those of the target language, and R as the mapping matrix. The solution to the orthogonal Procrustes problem is

$$BA^T = U\Sigma V^T, \tag{5}$$

$$R = UV^T, \tag{6}$$

where the first equation denotes singular value decomposition.

Besides solving orthogonal Procrustes problems, Xing et al. [10] propose to replace the singular values of the mapping matrix with one after every update of the matrix.

Weak orthogonal constraint

Conneau et al. [3] have applied an update rule to force the mapping matrix to be close to orthogonal matrix:

$$W \leftarrow (1+\beta)W - \beta(WW^T)W, \tag{7}$$

where β is set to be 0.01. This update rule makes all the eigenvectors of W have modulus close to 1.

Besides, in the studies of Zhang et al. [11] and Patra et al. [6], a loss for the orthogonality constraint is designed:

$$L = -\cos(x, W^T W x), \tag{8}$$

where x denotes an instance of source embeddings. The loss is jointly applied with other losses in training.

3 Proposed Model

As Patra et al. [6] argue that strictly orthogonal mapping is not optimal for bilingual word embedding alignment, we propose to apply approximately linear mapping matrices by minimizing the joint loss L_{total}:

$$L_{map} = \|WX - Y\|, \tag{9}$$

$$L_{orth} = \|WW^T - I\|, \tag{10}$$

$$L_{total} = \frac{\alpha}{\alpha+1} L_{map} + \frac{1}{\alpha+1} L_{orth}, \tag{11}$$

where X and Y are the word embeddings of source language words and target language words respectively in the lexicon, W is the mapping matrix, I is the identity matrix, and α is a hyperparameter. The L_{map} aims at maximizing the mapping of lexicon word embeddings. We apply the Euclidean distance as the mapping target to be aligned with the Procrustes problem, although applying the RCSLS loss proposed by Joulin et al. [4] will result in better BLI performance. L_{orth} puts a weak constraint on the orthogonality of the mapping matrix. It is not easy to control the orthogonality of the mapping matrix. We use the α to control the hardness of the orthogonality constraint. In later experiments, we find larger α results in more approximate orthogonality.

In the proposed method, the mapping matrix is approximately orthogonal, which makes the alignment more flexible while keeping much of the desirable features of orthogonal mapping.

Some related studies are optimizing the mapping matrices with the help of the loss concerning the orthogonality constraint, such as the work of Patra et al. [6]. But they do not control the orthogonality of the matrix, and they have not confirmed the exact amount of contribution to the improvement in mapping brought by approximate orthogonality.

4 Experiments

Experiment setting

The evaluation is based on the supervised BLI task of the MUSE dataset [1] [3]. The MUSE dataset consists of word embeddings trained by Bojanowski et al. [2] on Wikipedia and bilingual dictionaries generated by internal translation tools used at Facebook. We compare the results by the P@1 (precision of the first nearest neighbor lexicon induction) of BLI on CSLS nearest neighbor extraction.

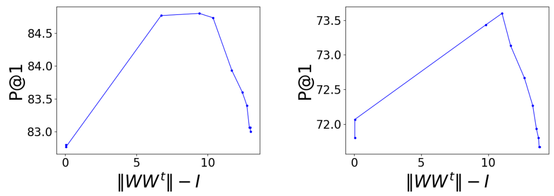
The results are compared with direct solutions to orthogonal Procrustes problems to check the pure improvement brought by approximate orthogonality. To show how powerful the approximately orthogonal mapping is, we also compare our results with those from MUSE(S) and MUSE(S) [3]. MUSE(S) learns the mapping under supervised settings by the iterative Procrustes. In the iterative Procrustes, the mapping is first solved from orthogonal Procrustes problem, and a new lexicon is generated by the mapping, and then a new mapping is solved from orthogonal Procrustes problem again with the generated lexicon. Such process is done multiple times. MUSE(U) optimizes the mapping matrix using GAN under unsupervised settings and apply iterative Procrustes refinement after the adversarial training.

[1] https://github.com/facebookresearch/MUSE

The hyperparameter α is tuned between 0.05 and 20, the parameters are updated with stochastic gradient descent, and the CSLS is applied in finding the nearest neighbors during lexicon generation.

Experiment results

We find that the degree of approximation in the mapping orthogonality influences the mapping performance. Two examples (es-en and de-en) are shown in Figure 1. Note that we indicate the degree of approximation in orthogonality by $\|WW^T - I\|$ (W is the optimized mapping matrix). The curves in Figure 1 imply that there are best degrees of orthogonality approximation for embedding mapping. For es-en, the best degree is 9.41; for de-en, that is 11.03.



(a) es-en (b) de-en

Figure 1: P@1 score for es-en (1a) and de-en (1b) when the mapping orthogonality varies.

The results for the whole MUSE dataset are shown in Table 1. From the table it is shown that approximately orthogonal mapping is superior to the orthogonal mapping obtained from solving the orthogonal Procrustes problem. Approximately orthogonal mapping also achieves higher precision than MUSE in most cases. For en-zh and zh-en where the two languages are very different, approximately orthogonal mapping behaves over absolute 11% better than orthogonal mapping.

The results indicate that due to the non-isomorphism of word embedding spaces, the best mapping should be approximately orthogonal mapping.

Approximate orthogonality refinement

Patra et al. [6] apply a weak orthogonality constraint on their model BLISS(R) which learns the mapping through GAN under semi-supervised settings. They also utilize iterative Procrustes refinement after training the approximately orthogonal mapping matrix. However, in our preliminary experiments of reproducing their experiments, it is found that the orthogonal refinement does not benefit the mapping performance (producing worse BLI precision). We propose to utilize approximate orthogonal refinement. In approximate orthogonal refinement, the generated lexicon is also used as the seed lexicon, but the new mapping matrix is optimized with our proposed approximately orthogonal mapping model. We test the approximate orthogonality refinement on the model of BLISS(R) on the MUSE dataset. Note that because BLISS(R) applies RCSLS as the mapping target, we also applies RCSLS as the target for refinement. The result is shown in Table 2.

As expected, the approximate orthogonality refinement is better than Procrustes refinement and improves the BLI performance in some cases. The performance of approximate orthogonality refinement degrades for en-zh, probably for the generated lexicon is not satisfactory.

5 Conclusion

We analyze the effect of approximately orthogonal mapping for BLI. The approximately orthogonal mapping outperforms orthogonal mapping through solving the orthogonal Procrustes problems and also outperforms MUSE(S) and MUSE(U) in most cases. The benefit of approximate orthogonality especially helps mapping between English and traditional Chinese which are a linguistically distant language pair.

Based on the analysis, we propose approximate orthogonality refinement which boosts the BLI performance of BLISS(R).

While many researchers tend to follow the strict orthogonality constraint for bilingual word embedding alignment, our experimental results suggest researchers to always try approximate orthogonality.

References

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

[2] 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.

[3] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *Proceedings of ICLR*, 2018.

[4] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Table 1: Bilingual lexicon induction P@1 for MUSE dataset. The improvement denotes how much the approximate orthogonality performs better than vanilla Procrustes. Note that there are no results for en-eo and eo-en in the table, for we have not found the data for eo (seems to be deleted). The results of MUSE for en-zh and zh-en is different from the results written in the paper of MUSE, maybe for the dataset has been updated.

Method	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en
Procrustes	81.40	82.87	81.07	82.40	73.47	72.40	51.67	63.67	32.47	25.13
MUSE(S)	81.87	83.47	**82.13**	82.40	74.27	72.73	51.67	63.67	32.47	25.13
MUSE(U)	81.20	83.33	81.53	82.53	**74.87**	73.47	35.40	59.80	0.00	0.00
Ours	**82.07**	**84.80**	81.73	**83.26**	73.83	**73.60**	**53.40**	**64.00**	**43.60**	**40.20**
improvement	0.67	1.93	0.66	0.86	0.36	1.20	1.73	0.33	11.13	15.07
$\|WW^T - I\|$	7.38	9.41	6.96	9.59	8.82	11.03	12.36	11.23	11.65	2.83

Table 2: Bilingual lexicon induction P@1 by BLISS and refinement.

Method	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en
BLISS(R)	83.60	86.47	84.20	**84.73**	**78.33**	**76.33**	57.07	**67.07**	**48.33**	**47.93**
Procrustes Refinement	82.87	84.33	82.93	83.60	76.93	72.87	54.47	65.13	36.40	26.27
Approximate Orthogonality Refinement	**83.80**	**86.73**	**84.33**	84.60	77.27	75.60	**57.27**	66.0	36.40	45.73

Processing, pages 2979–2984, Brussels, Belgium, 2018. Association for Computational Linguistics.

[5] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4, 2013.

[6] Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy, 2019. Association for Computational Linguistics.

[7] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531, 2010.

[8] Samuel L Smith, David H P Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *5th International Conference on Learning Representations (ICLR 2017)*, 2017.

[9] Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, 2018. Association for Computational Linguistics.

[10] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proceedings of the 2015 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado, 2015. Association for Computational Linguistics.

[11] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial Training for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, jul 2017. Association for Computational Linguistics.

[12] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Earth Mover{'}s Distance Minimization for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark, sep 2017. Association for Computational Linguistics.