A Multi-Pairwise Extension of Procrustes Analysis for Multilingual Word Translation

Hagai Taitelbaum
Faculty of Engineering
Bar-Ilan University, Israel
hagait62@gmail.com

Gal Chechik
The Gonda Brain Research Center
Bar-Ilan University, NVIDIA
gal.chechik@biu.ac.il

Jacob Goldberger
Faculty of Engineering
Bar-Ilan University, Israel
jacob.goldberger@biu.ac.il

Abstract

In this paper we present a novel approach to simultaneously representing multiple languages in a common space. Procrustes Analysis (PA) is commonly used to find the optimal orthogonal word mapping in the bilingual case. The proposed *Multi Pairwise Procrustes Analysis* (MPPA) is a natural extension of the PA algorithm to multilingual word mapping. Unlike previous PA extensions that require a k-way dictionary, this approach requires only pairwise bilingual dictionaries that are much easier to construct in either a supervised or an unsupervised way. The improved performance of the MPPA algorithm is demonstrated on two standard multilingual tasks.

1 Introduction

Continuous word embeddings have been proved effective in numerous NLP applications. In cross-language tasks, these vector-space representations have recently emerged as a tool to transfer knowledge from one language to another. Specifically, several studies have suggested forming cross-lingual embeddings by learning a linear mapping from a source-language embedding space to a target-language one and demonstrated the benefits of this approach for word translation (Mikolov et al., 2013; Klementiev et al., 2012). Xing et al. (2015) showed that imposing orthogonality constraints on the linear mapping between spaces can alleviate overfitting. Building on these concepts, several studies have aimed to improve these bilingual word embeddings using bilingual word dictionaries that are created in either a supervised or an unsupervised manner (Artetxe et al., 2017a). Bilingual word embedding were found to be useful in a number of monolingual and cross-lingual NLP tasks (Vulic and Moens, 2015; Tsai and Roth, 2016).

Bilingual embedding can be extended to a multilingual setup by jointly learning mappings from each monolingual word embedding to a shared vector space. Modeling multiple languages jointly has been shown to improve modeling accuracy on bilingual tasks because it can utilize knowledge learned from the other languages (Ammar et al., 2016; Duong et al., 2017; Taitelbaum et al., 2019).

Extending the bilingual setup to a multilingual setting poses new challenges. For bilingual embedding, the word-mapping problem has a closed-form solution known as Orthogonal Procrustes Analysis (PA), which can be computed using singular value decomposition (Schnemann, 1966). However, there is no similar closed-form solution for the multi-language case. The standard extension of PA to multi-set alignment is Generalized Procrustes Analysis (GPA) (Gower, 1975) which is an iterative greedy algorithm. GPA was recently used to jointly transform multiple languages into a shared vector space (Kementchedjhieva et al., 2018). However, GPA assumes that a multi-way word correspondence is available, which is often not the case. Building a multi-way dictionary is a challenging task in itself.

In this study, we propose a novel efficient approach for mapping multiple languages simultaneously into a shared vector space, while enforcing orthogonality constraints. This approach, *Multi Pairwise Procrustes Analysis* (MPPA) can be viewed as a multilingual extension of the Procrustes Analysis. Unlike GPA-based approaches, MPPA does not require a multi-way dictionary, but only bilingual dictionaries which are much easier to obtain even in an unsupervised manner. We evaluated MPPA on two standard multilingual tasks and report better results than GPA based methods and competitive results with gradient based methods.

Our main contribution is a new, efficient, and

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3560–3565,
Hong Kong, China, November 3–7, 2019.

easy-to-use algorithm for solving the extension of the Orthogonal Procrustes problem to the multilingual case. Our project code will be publicly available.

2 A Multi-Pairwise Extension of Procrustes Analysis

We first briefly review Procrustes Analysis (PA), a procedure to find the best orthogonal mapping between two languages. We then describe our approach, *Multi-Pairwise Procrustes Analysis* (MPPA), which extends PA to the multilingual case.

Assume we are given d-dimensional word embedding data from two languages along with a dictionary consisting of pairs of corresponding words. Mikolov et al. (2013) showed that there is a strong linear correlation between the vector spaces of two languages and that learning a complex non-linear neural mapping does not yield better results than with a linear mapping. Xing et al. (2015) further showed that enforcing the linear mappings to be orthogonal matrices reduces overfitting and improves performance. We can learn the orthogonal mapping T by minimizing the following cost function:

$$S(T) = \sum_{t=1}^{n} \|Tx_t - y_t\|^2 , \qquad (1)$$

where x_t and y_t are embeddings pf corresponding words from the two languages and n is the dictionary size. Schnemann (1966) proved that the solution to Eq. (1), obtained as the result of a Procrustes Analysis algorithm, is $T = UV^\top$, where $U\Sigma V^\top$ is the singular value decomposition (SVD) of the $d \times d$ matrix $M = \sum_t y_t x_t^\top$. This method has been used in many recent cross-lingual studies (Xing et al., 2015; Artetxe et al., 2016, 2017a,b, 2018a,b; Hamilton et al., 2016; Conneau et al., 2017; Ruder et al., 2018).

Assume we are given d-dimensional word embedding data from k languages and that each pair of languages is provided with a dictionary composed of pairs of corresponding words from the two languages. We could learn a mapping for each language pair independently as a solution to Eq. (1). However, this approach does not benefit from the multilingual setup. Another approach would be to choose one of the languages as a "pivot" and learn a mapping from each language to the pivot separately. A typical choice for the pivot, used in publicly available aligned vectors, is English (Conneau et al., 2017; Joulin et al., 2018). This strategy, however, does not guarantee that the indirect word translation between language pairs will have high quality. Alternatively, we can enforce a transitivity constraint by mapping all the embedding spaces to a shared vector space. Our goal in the multilingual case is thus to find the orthogonal matrices $T_1, ..., T_k$ such that pairs of corresponding words from different languages are mapped into close vectors in the shared space. Formally, we want to minimize the following mean-square error score:

$$S(T_1, ..., T_k) = \sum_{i<j} \sum_{t=1}^{n_{ij}} \|T_i x_{ij,t} - T_j x_{ji,t}\|^2 \quad (2)$$

where $(x_{ij,t}, x_{ji,t})$ is a pair of corresponding words in the i and j languages, respectively and n_{ij} is the dictionary size. We use this notation to emphasize that the vocabularies of the same language in different dictionaries are not necessarily the same.

When more than two languages are involved there is no closed-form solution to the global minimum of (2). We propose an efficient algorithm for minimizing it. The basic step is optimizing the score (2) with respect to the mapping T_i while keeping all other mappings fixed. Viewing the objective score (2) as a function of T_i we obtain:

$$S(T_i) = \sum_{j \neq i} \sum_{t=1}^{n_{ij}} \|T_i x_{ij,t} - y_{j,t}\|^2 + \text{const.} \quad (3)$$

where $y_{j,t} = T_j x_{ji,t}$ is the representation of $x_{ji,t}$ in the common space. This is exactly the Orthogonal Procrustes problem (1) of finding a mapping from language i into the common space. The optimal orthogonal matrix T_i is thus obtained by $T_i = UV^\top$ s.t. $U\Sigma V^\top$ is the SVD of the matrix:

$$M_i = \sum_{j \neq i} \sum_{t} T_j x_{ji,t} x_{ij,t}^\top . \qquad (4)$$

Once T_i is updated we move to the next language in a circular manner. At each step in the iterative algorithm, the score (2) is monotonically decreased until it converges to a local minimum point. Hence, we can stop the optimization procedure once there is no significant improvement in the objective score (2).

Each iteration is very costly since we need to go over all the dictionary words. To avoid this,

we can compute cross correlation $d \times d$ matrix for each pair of languages i, j in a preprocessing step:

$$C_{ij} = C_{ji}^{\top} = \sum_{t} x_{ji,t} x_{ij,t}^{\top}. \quad (5)$$

Substituting (5) in (4) we obtain that

$$M_i = \sum_{j \neq i} T_j C_{ij}. \quad (6)$$

Therefore, updating the mapping T_i can be done in a very efficient way without going over all the bilingual dictionaries of the i-th language.

Algorithm 1 Multi Pairwise Procrustes Analysis

Required: A set of lexical of word pairs between each pair of languages.

Task: Find a set of orthogonal mappings $T_1, .., T_k$, to a common space.

Compute cross-correlation matrices:

$$C_{ij} = C_{ji}^{\top} = \sum_{t} x_{ji,t} x_{ij,t}^{\top} \qquad 1 \leq i < j \leq k$$

Initialization: $T_1 = I$,
for $i = 2, ..., k$ **do**
  $U\Sigma V^{\top}$ = SVD($\sum_{j<i} T_j C_{ij}$)
  $T_i \leftarrow UV^{\top}$
end
Algorithm:
while *not converged* **do**
  for $i = 1, ..., k$ **do**
    $U\Sigma V^{\top}$ = SVD($\sum_{j \neq i} T_j C_{ij}$)
    $T_i \leftarrow UV^{\top}$
  end
end

The proposed Multi Pairwise Procrustes Analysis (MPPA) word mapping training procedure is depicted in Algorithm box 1. The algorithm description also contains an initialization procedure that can help avoid getting stuck at local optima. The idea of the initialization is aligning each new language i to the current common space which was built with languages $j < i$.

MPPA requires only pairwise bilingual dictionaries. It is applicable even if we only have dictionaries for a subset of all the language pairs such that each language under consideration is represented in at least one bilingual dictionary. Consider a graph whose vertices are the languages and an edge indicates the existence of a dictionary between two languages. It can easily be seen that if the graph is loop-free (as in the case where we only have dictionaries for a pivot language) the optimization of (2) is decoupled and each bilingual mapping can be learned separately. The task become really multi-lingual once the graph is loopy, where mapping transitivity implies that there is more than a single path between the source and target languages. We note in passing that we can consider the word representation in the common space as a latent variable and the mapping matrices as unknown parameters. The MPPA algorithm can be thus viewed as an instance of the EM algorithm (Dempster et al., 1977). Further discussion regarding the connection between MPPA algorithm and the EM algorithm can be found in (Goldberger, 1999).

	he	af	oc	et	bs	Avg.
PA	37.5	28.9	17.1	30.0	22.4	27.2
MGPA	37.5	28.9	23.8	30.7	21.0	28.4
MPPA	40.2	32.1	25.4	35.5	26.2	**31.9**

Table 1: p@1 for low resource languages: Hebrew, Afrikaans, Occitan, Estonian, and Bosnian, trained with multilingual algorithms over triplets.

3 Related work

The standard extension of PA to multi-set alignment is *Generalized Procrustes Analysis* (GPA) (Gower, 1975). Kementchedjhieva et al. (2018) recently proposed the Multi-support GPA (MGPA) algorithm for multilingual word translation which is based on the GPA. Their algorithm requires a k-way dictionary in the form of (x_{it}) where $(x_{1t}, ..., x_{kt})$ are representations of words that share the same semantic meaning across all the k languages. This multi-way dictionary is constructed from the bilingual dictionaries (Kementchedjhieva et al., 2018). Whereas conflating multiple senses of a word is already problematic for bilingual dictionaries, this issue is amplified in a multilingual vocabulary. In our approach we avoid this form of error-prone data processing that consists of finding a joint translation of a single word across all the languages. Instead, the MPPA algorithm uses the bilingual dictionaries directly. Note that MPPA is an extension of the GPA algorithm. In case we are given a multi-way dictionary GPA and MPPA optimize the same cost function and MPPA can be viewed as an efficient alterna-

	en-de	en-fr	en-es	en-it	en-pt	de-en	de-fr	de-es	de-it	de-pt	fr-en	fr-de	fr-es	fr-it	fr-pt	
PA	73.5	81.1	81.4	77.3	79.9	72.4	73.3	67.7	69.5	59.1	82.4	69.5	82.6	83.2	78.1	
MAT+MPPA	74.5	82.7	82.2	78.5	81.3	72.9	75.2	68.0	70.1	61.1	82.2	69.0	83.6	83.1	78.7	
MAT+MPSR	74.8	82.4	82.5	78.8	81.5	72.9	76.7	69.6	72.0	63.2	81.8	71.2	83.9	83.5	79.3	
UMH	75.1	82.7	82.5	78.9	82.0	75.5	73.5	67.2	68.7	59.0	83.1	69.8	82.7	82.5	77.5	
	es-en	es-de	es-fr	es-it	es-pt	it-en	it-de	it-fr	it-es	it-pt	pt-en	pt-de	pt-fr	pt-es	pt-it	Avg.
PA	82.9	68.3	85.8	83.5	87.3	76.9	67.5	87.1	87.3	81.0	80.3	63.7	84.3	91.5	81.1	78.0
MAT+MPPA	83.5	66.5	85.9	83.7	86.8	77.7	67.1	87.7	87.5	81.2	80.2	63.7	84.6	92.2	82.6	78.5
MAT+MPSR	83.7	69.0	86.9	84.5	87.8	77.4	69.5	88.1	88.2	82.3	79.9	65.7	86.3	92.7	82.6	**79.3**
UMH	85.3	68.7	85.1	83.3	86.3	79.9	67.5	86.7	87.0	80.4	82.1	64.4	83.6	91.7	81.1	78.5

Table 2: Multilingual word translation results for English, German, French, Spanish, Italian and Portuguese. The reported numbers are precision@1 in percentage.

tive to the GPA optimization procedure.

Another line of research applies stochastic gradient-based optimization methods to minimize the mean-square error score (2) jointly with refinement of the bilingual dictionaries. The gradient is approximated by sampling word pairs from the bilingual dictionaries. Chen and Cardie (2018) proposed the *Multilingual Pseudo Supervised Refinement* (MPSR) for this minimization task that uses simple gradient methods in order to minimize (2). For unsupervised setup Chen and Cardie (2018) used an adversarial initialization step, *Multilingual Adversarial Training* (MAT). Alaux et al. (2019) presented, *Unsupervised Multilingual Hyperalignment* (UMH), a similar algorithm that extends the bilingual methods proposed by Grave et al. (2018); Alvarez-Melis and Jaakkola (2018), to multilingual setup.

A main difference between UMH (Alaux et al., 2019) and MAT+MPSR (Chen and Cardie, 2018) is how they treat orthogonality. The first is a stochastic gradient optimization followed by a projection on the set of orthogonal matrices. In the second method orthogonality is a regularization term that is optimized by gradient methods. The matrices are encouraged to be orthogonal by an orthogonalization update (Cisse et al., 2017) that yields matrices that are close to orthogonal but are not necessarily exactly orthogonal. In contrast to gradient based methods, our approach avoids word sampling and hyper-parameters that need to be tuned.

4 Experiments

Datasets and embeddings We used the MUSE benchmark (Conneau et al., 2017)[1], which consist of bilingual dictionaries of 5000 unique source word for training and 1500 for testing. The fastText embeddings (Bojanowski et al., 2017) trained on Wikipedia data, are available online[2]. Vectors were normalized to unit length and then zero centered (Artetxe et al., 2016).

Compared methods We compared MPPA to MGPA (Kementchedjhieva et al., 2018), MAT+MPSR (Chen and Cardie, 2018) and UMH (Alaux et al., 2019). We used the task and results reported in the corresponding paper (UMH results are from the appendix). All methods ran several refinement epochs (Artetxe et al., 2017a), where after each refinement iteration dictionaries were re-build, as described in Conneau et al. (2017). Model selection was done by the best validation criterion suggested in Conneau et al. (2017) and extended in Chen and Cardie (2018). All these methods retrieve word translation using the *Cross-domain Similarity Local Scaling* (CSLS) criterion (Lample et al., 2018).

Results The first experiment was conducted over language triplets (Kementchedjhieva et al., 2018). The goal is to translate from English to a low resource language (like Bosnian) using a high resource language (like Russian). As in Kementchedjhieva et al. (2018), 10 refinement epochs were used, and initial dictionaries for each language pair were generated by pairs of words with identical string matching.

Table 1 depicts precision@1 for the triplets task. MPPA outperformed MGPA and both outperform PA. Note that MGPA needs a multi-way dictionary constructed from the bilingual dictionaries. In contrast, MPPA uses directly the raw data (the bilingual dictionaries).

The second experiment involved multilingual word translation in six European languages: English, German, French, Spanish, Italian and Por-

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

[2] https://github.com/facebookresearch/fastText

tuguese (Lample et al., 2018). We compared MPPA to MAT+MPSR (Chen and Cardie, 2018). MAT+MPSR is an unsupervised method, so for a fair comparison we replaced the MPSR algorithm with our MPPA algorithm, thus obtaining MAT+MPPA. We ran 5 refinement epochs, after the MAT step, as the default option in MAT+MPSR source code[3]. MPPA training phase is 10 times faster than MPSR equivalent phase, which also have hyper-parameters that needed to be tuned. UMH (Alaux et al., 2019), was also evaluated on this benchmark.

Table 2 shows precision@1 results. MPPA was comparable to UMH and MPSR performed slightly better. Note that the MPSR mapping matrices were not exactly orthogonal. They indeed achieved smaller mean-square error (2) on the training data than our solution, which was restricted to be orthogonal. This suggests that the orthogonality constraint, especially in the multilingual case where it is combined with transitivity constraints, can be too restrictive.

5 Conclusion

This paper presents a general approach to map word embeddings into a common space that can be viewed as an extension of PA to the multilingual case. The proposed algorithm efficiently avoids the need to go over the whole dictionary at each iteration. The optimization is done by enforcing both transitivity and orthogonal constraints. A possible future research direction would involve finding efficient optimization methods where the orthogonality constraint could be slightly relaxed.

References

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyperalignment for multilingual word embeddings. In *International Conference on Learning Representations*.

David Alvarez-Melis and Tommi S Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*.

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

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

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

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 Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017b. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

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.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. 2017. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 854–863. JMLR. org.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39:1–38.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Coh. 2017. Multilingual training of crosslingual word embeddings. In *The Conference of the European Chapter of the Association for Computational Linguistics*.

Jacob Goldberger. 1999. Registration of multiple point sets using the EM algorithm. In *IEEE International Conference on Computer Vision*.

John C Gower. 1975. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

[3] https://github.com/ccsasuke/umwe

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with wasserstein procrustes. *CoRR*, abs/1805.11222.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Herve Jegou, , and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing procrustes analysis for better bilingual dictionary induction. In *CONLL*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474.

Guillaume Lample, Alexis Conneau, Marc-Aurelio Ranzato, Ludovic Denoyer, and Herv Jgou. 2018. Word translation without parallel data. In *International Conference on Learning Representation*.

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

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. 2018. A discriminative latent-variable model for bilingual lexicon induction. *arXiv preprint arXiv:1808.09334*.

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

Hagai Taitelbaum, Gal Chechik, and Jacob Goldberger. 2019. Multilingual word translation using auxiliary languages. In *Conference on Empirical Methods in Natural Language Processing*.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *NAACL-HLT*.

Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embedding. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, page 363372.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.