Generalizing Procrustes Analysis for Better Bilingual Dictionary Induction

Yova Kementchedjhieva[◇] **Sebastian Ruder**[♠♣] **Ryan Cotterell**[♡] **Anders Søgaard**[◇]
[◇]University of Copenhagen, Copenhagen, Denmark
[♠]Insight Research Centre, National University of Ireland, Galway, Ireland
[♣]Aylien Ltd., Dublin, Ireland
[♡]University of Cambridge, Cambridge, UK
{yova|soegaard}@di.ku.dk,sebastian@ruder.io,ryan.cotterell@gmail.com,

Abstract

Most recent approaches to bilingual dictionary induction find a linear alignment between the word vector spaces of two languages. We show that projecting the two languages onto a third, latent space, rather than directly onto each other, while equivalent in terms of expressivity, makes it easier to learn approximate alignments. Our modified approach also allows for supporting languages to be included in the alignment process, to obtain an even better performance in low resource settings.

1 Introduction

Several papers recently demonstrated the potential of very weakly supervised or entirely unsupervised approaches to bilingual dictionary induction (BDI) (Barone, 2016; Artetxe et al., 2017; Zhang et al., 2017; Conneau et al., 2018; Søgaard et al., 2018), the task of identifying translational equivalents across two languages. These approaches cast BDI as a problem of aligning monolingual word embeddings. Pairs of monolingual word vector spaces can be aligned without any explicit cross-lingual supervision, solely based on their distributional properties (for an adversarial approach, see Conneau et al. (2018)). Alternatively, weak supervision can be provided in the form of numerals (Artetxe et al., 2017) or identically spelled words (Søgaard et al., 2018). Successful unsupervised or weakly supervised alignment of word vector spaces would remove much of the data bottleneck for machine translation and push horizons for cross-lingual learning (Ruder et al., 2018).

In addition to an unsupervised approach to aligning monolingual word embedding spaces with adversarial training, Conneau et al. (2018) present a supervised alignment algorithm that assumes a gold-standard seed dictionary and performs Procrustes Analysis (Schönemann, 1966). Søgaard et al. (2018) show that this approach, weakly supervised with a dictionary seed of *cross-lingual homographs*, i.e. words with identical spelling across source and target language, is superior to the completely unsupervised approach. We therefore focus on weakly-supervised Procrustes Analysis (PA) for BDI here.

The implementation of PA in Conneau et al. (2018) yields notable improvements over earlier work on BDI, even though it learns a simple linear transform of the source language space into the target language space. Seminal work in supervised alignment of word vector spaces indeed reported superior performance with linear models as compared to non-linear neural approaches (Mikolov et al., 2013). The relative success of the simple linear approach can be explained in terms of isomorphism across monolingual semantic spaces,[1] an idea that receives support from cognitive science (Youn et al., 1999). Word vector spaces are not *perfectly* isomorphic, however, as shown by Søgaard et al. (2018), who use a Laplacian graph similarity metric to measure this property. In this work, we show that projecting both source and target vector spaces into a *third* space (Faruqui and Dyer, 2014), using a variant of PA known as Generalized Procrustes Analysis (Gower, 1975), makes it easier to learn the alignment between two word vector spaces, as compared to the single linear transform used in Conneau et al. (2018).

Contributions We show that Generalized Procrustes Analysis (GPA) (Gower, 1975), a method that maps two vector spaces into a third, latent space, is superior to PA for BDI, e.g., improving the state-of-the-art on the widely used English-Italian dataset (Dinu et al., 2015) from a P@1 score of 66.2% to 67.6%. We compare GPA to PA

[1]Two vector spaces are isomorphic if there is an invertible linear transformation from one to the other.

on aligning English with five languages representing different language families (Arabic, German, Spanish, Finnish, and Russian), showing that GPA consistently outperforms PA. GPA also allows for the use of additional support languages, aligning three or more languages at a time, which can boost performance even further. We present experiments with multi-source GPA on an additional five low-resource languages from the same language families (Hebrew, Afrikaans, Occitan, Estonian, and Bosnian), using their bigger counterpart as a support language. Our code is publicly available.[2]

2 Procrustes Analysis

Procrustes Analysis is a graph matching algorithm, used in most mapping-based approaches to BDI (Ruder et al., 2018). Given two graphs, E and F, Procrustes finds the linear transformation T that minimizes the following objective:

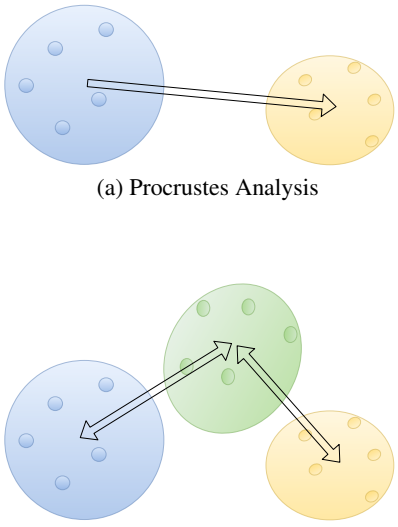
$$\arg\min_{T} ||TE - F||^2 \quad (1)$$

thus minimizing the trace between each two corresponding rows of the transformed space TE and F. We build E and F based on a seed dictionary of N entries, such that each pair of corresponding rows in E and F, (e_n, f_n) for $n = 1, \ldots, N$ consists of the embeddings of a translational pair of words. In order to preserve the monolingual quality of the transformed embeddings, it is beneficial to use an orthogonal matrix T for cross-lingual mapping purposes (Xing et al., 2015; Artetxe et al., 2017).[3] Conveniently, the orthogonal Procrustes problem has an analytical solution, based on Singular Value Decomposition (SVD):

$$\begin{aligned} F^\top E &= U\Sigma V^\top \\ T &= VU^\top \end{aligned} \quad (2)$$

3 Generalized Procrustes Analysis

Generalized Procrustes Analysis (Gower, 1975) is a natural extension of PA that aligns k vector spaces at a time. Given embedding spaces



(a) Procrustes Analysis

(b) Generalized Procrustes Analysis

Figure 1: Visualization of the difference between PA, which maps the source space directly onto the target space, and GPA, which aligns both source and target spaces with a third, latent space, constructed by averaging over the two language spaces.

$E_1, \ldots, E_k$, GPA minimizes the following objective:

$$\arg\min_{\{T_1,\ldots,T_k\}} \sum_{i<j}^{k} ||T_i E_i - T_j E_j)||^2 \quad (3)$$

For an analytical solution to GPA, we compute the average of the embedding matrices $E_{1\ldots k}$ after transformation by $T_{1\ldots k}$:

$$G = k^{-1} \sum_{i=1}^{k} E_i T_i \quad (4)$$

thus obtaining a latent space, G, which captures properties of each of $E_{1\ldots k}$, and potentially additional properties emerging from the combination of the spaces. On the very first iteration, prior to having any estimates of $T_{1\ldots k}$, we set $G = E_i$ for a random i. The new values of $T_{1\ldots k}$ are then obtained as:

$$\begin{aligned} G^\top E_i &= U\Sigma V^\top \\ T_i &= VU^\top \text{ for } i \text{ in } 1 \ldots k \end{aligned} \quad (5)$$

Since G is dependent on $T_{1\ldots k}$ (see Eq.4), the solution of GPA cannot be obtained in a single step (as is the case with PA), but rather requires that we loop over subsequent updates of G (Eq.4) and $T_{1\ldots k}$ (Eq.5) for a fixed number of steps or until

[2] https://github.com/YovaKem/generalized-procrustes-MUSE

[3] Recently, Doval et al. (2018) showed that the monolingual quality of embeddings need not suffer from a transformation guided by cross-lingual alignment, but their method still relies on an initial alignment obtained e.g. with Procrustes analysis, as described here.

High-resource	AR	DE	ES	FI	RU
	575k	2,183k	1,412k	437k	1,474k
Low-resource	HE	AF	OC	ET	BS
	224k	49k	84k	175k	77k

Table 1: Statistics for Wikipedia corpora.

satisfactory convergence. We observed little improvement when performing more than 100 updates, so we fixed that as the number of updates.

Notice that for $k = 2$ and with the orthogonality constraint in place, the objective for Generalized Procrustes Analysis (Eq. 3) reduces to that for simple Procrustes (Eq. 1):

$$\begin{aligned} &\arg \min_{\{T_1,T_2\}} ||T_1E_1 - T_2E_2||^2 \\ &= \arg \min_{T} ||TE_1 - E_2||^2 \\ &\text{where } T = T_1T_2^T \end{aligned} \tag{6}$$

Here T itself is also orthogonal. Yet, the solution found with GPA may differ from the one found with simple Procrustes: the former maps E_1 and E_2 onto a third space, G, which is the average of the two spaces, instead of mapping E_1 directly onto E_2. To understand the consequences of this difference, consider a single step of the GPA algorithm where after updating G according to Eq.4 we are recomputing T_1 using SVD. Due to the fact that G is partly based on E_1, these two spaces are bound to be more similar to each other than E_1 and E_2 are.[4] Finding a good mapping between E_1 and G, i.e. a good setting of T_1, should therefore be easier than finding a good mapping from E_1 to E_2 directly. In this sense, by mapping E_1 onto G, rather than onto E_2 (as PA would do), we are solving an easier problem and reducing the chance of a poor solution.

4 Experiments

In our experiments, we generally use the same hyper-parameters as used in Conneau et al. (2018), unless otherwise stated. When extracting dictionaries for the bootstrapping procedure, we use cross-domain local scaling (CSLS, see Conneau et al. (2018) for details) as a metric for ranking candidate translation pairs, and we only use the ones that rank higher than 15,000. We do not put any restrictions on the initial seed dictionaries, based on cross-lingual homographs: those vary considerably in size, from 17,012 for Hebrew to 85,912 for Spanish. Instead of doing a single training epoch, however, we run PA and GPA with early stopping, until five epochs of no improvement in the validation criterion as used in Conneau et al. (2018), i.e. the average cosine similarity between the top 10,000 most frequent words in the source language and their candidate translations as induced with CSLS. Our metric is Precision at k×100 (P@k), i.e. percentage of correct translations retrieved among the k nearest neighbor of the source words in the test set (Conneau et al., 2018). Unless stated otherwise, experiments were carried out using the publicly available pre-trained fastText embeddings, trained on Wikipedia data,[5] and bilingual dictionaries—consisting of 5000 and 1500 unique word pairs for training and testing, respectively—provided by Conneau et al. (2018)[6].

4.1 Comparison of PA and GPA

High resource setting We first present a direct comparison of PA and GPA on BDI from English to five fairly high-resource languages: Arabic, Finnish, German, Russian, and Spanish. The Wikipedia corpus sizes for these languages are reported in Table 1. **Results** are listed in Table 2. GPA improves over PA consistently for all five languages. Most notably, for Finnish it scores 2.5% higher than PA.

Common benchmarks For a more extensive comparison with previous work, we include results on English–{Finnish, German, Italian} dictionaries used in Conneau et al. (2018) and Artetxe et al. (2018)—the second best approach to BDI known to us, which also uses Procrustes Analysis. We conduct experiments using three forms of supervision: gold-standard seed dictionaries of 5000 word pairs, cross-lingual homographs, and numerals. We use train and test bilingual dictionaries from Dinu et al. (2015) for English-Italian and from Artetxe et al. (2017) for English-{Finnish, German}. Following Conneau et al. (2018), we report results with a set of CBOW embeddings trained on the WaCky corpus (Barone, 2016), and with Wikipedia embeddings.

Results are reported in Table 3. We observe that

[4] A theoretical exception being the case there E_1 and E_2 are identical.

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

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

	AR		DE		ES		FI		RU		Ave	
	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$
PA	34.73	61.87	73.67	91.73	81.67	92.93	45.33	75.53	47.00	79.00	56.48	80.21
GPA	**35.33**	**64.27**	**74.40**	**91.93**	**81.93**	**93.53**	**47.87**	**76.87**	**48.27**	**79.13**	**57.56**	**81.15**

Table 2: Bilingual dictionary induction performance, measured in P@k, of PA and GPA across five language pairs.

	IT			DE			FI		
	5000	Identical	Numerals	5000	Identical	Numerals	5000	Identical	Numerals
	WACKY								
Artetxe et al. (2018)	45.27*	38.33	39.40*	44.27*	40.73	40.27*	32.94*	27.39	26.47*
PA	44.90	45.47	01.13	47.26	47.20	45.93	**33.50**	**31.46**	01.05
GPA	**45.33**	**45.80**	**45.93**	**48.46**	**47.60**	**47.60**	31.39	31.04	**28.93**
	WIKIPEDIA								
PA	66.24	66.39	-	65.33	64.77	-	36.77	35.40	-
GPA	**67.60**	**67.14**	-	**66.21**	**65.81**	-	**38.14**	**37.87**	-

Table 3: Results on standard benchmarks, measured in P@1. * Results as reported in the original paper. **Notes:** Conneau et al. (2018) report 63.7 on Italian with Wikipedia embeddings; results with different embedding sets are not comparable due to a non-zero out-of-vocabulary rate on the test set for Wikipedia embeddings; Wikipedia embeddings are trained on corpora with removed numerals, so supervision from numerals cannot be applied.

GPA outperforms PA consistently on Italian and German with the WaCky embeddings, and on all languages with the Wikipedia embeddings. Notice that once more, Finnish benefits the most from a switch to GPA in the Wikipedia embeddings setting, but it is also the only language to suffer from that switch in the WaCky setup.

Interestingly, PA fails to learn a good alignment for Italian and Finnish when supervised with numerals, while GPA performs comparably with numerals as with other forms of supervision. Conneau et al. (2018) point out that improvement from subsequent iterations of PA is generally negligible, which we also found to be the case. We also found that while PA learned a slightly poorer alignment than GPA, it did so faster. With our criterion for early stopping, PA converged in 5 to 10 epochs, while GPA did so within 10 to 15 epochs[7] . In the case of Italian and Finnish alignment supervised by numerals, PA converged in 8 and 5 epochs, respectively, but clearly got stuck in local minima. GPA took considerably longer to converge: 27 and 74 epochs, respectively, but also managed to find a reasonable alignment between the language spaces. This points to an important difference in the learning properties of PA and GPA—unlike PA, GPA has a two-fold objective of opposing forces: it is simultaneously aligning each embedding space to two others, thus pulling it in different directions. This characteristic helps GPA avoid particularly adverse local minima.

4.2 Multi-support GPA

In these experiments, we perform GPA with $k = 3$, including a third, linguistically-related supporting language in the alignment process. To best evaluate the benefits of the multi-support setup, we use as targets five low-resource languages: Afrikaans, Bosnian, Estonian, Hebrew and Occitan (see statistics in Table 1)[8]. Three-way dictionaries, both the initial one (consisting of cross-lingual homographs) and subsequent ones, are obtained by assuming transitivity between two-way dictionaries: if two pairs of words, e^m–e^n and e^m–e^l, are deemed translational pairs, then we consider e^n–e^m–e^l a translational triple.

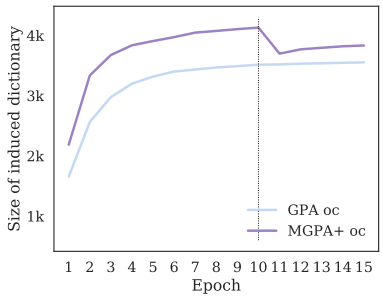
We report **results** in Table 4 with multi-support GPA in two settings: a three-way alignment trained for 10 epochs (MGPA), and a three-way alignment trained for 10 epochs, followed by 5

[7]Notice that one epoch with both PA and GPA takes less than half a minute, so the slower convergence of GPA is in no way prohibitive.

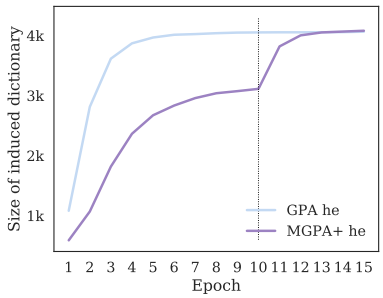
[8]Occitan dictionaries were not available from the MUSE project, so we extracted a test dictionary of 911 unique word pairs from an English-Occitan lexicon available at http://www.occitania.online.fr/aqui.comenca.occitania/en-oc.html.

	AF		BS		ET		HE		OC		Ave	
	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$	$k=1$	$k=10$
PA	28.87	50.53	22.40	48.40	30.00	57.93	37.53	67.27	17.12	33.26	27.18	51.48
GPA	**29.93**	**50.67**	**24.20**	**50.20**	**31.87**	**60.07**	38.93	**68.93**	17.12	34.91	28.41	52.96
MGPA	28.93	49.20	21.00	48.60	30.73	59.53	37.53	66.47	**23.82**	**40.18**	28.40	52.80
MGPA$^+$	28.80	49.20	23.46	48.87	31.27	59.80	**40.40**	68.80	22.83	38.53	**29.35**	**53.04**

Table 4: Results for low-resource languages with PA, GPA and two multi-support settings.



(a) Occitan



(b) Hebrew

Figure 2: Progression of dictionary size during GPA and MGPA+ training. The dotted line marks the boundary between MGPA and fine-tuning.

epochs of two-way fine-tuning (MGPA+). We observe that at least one of our new methods always improves over PA. GPA always outperforms PA and it also outperforms the multi-support settings on three out of five languages. Yet, results for Hebrew and especially for Occitan, are best in a multi-support setting—we thus mostly focus on these two languages in the following subsections.

MGPA has variable performance: for four languages precision suffers from the addition of a third language, e.g. compare 38.93 for Hebrew with GPA to 37.53 with MGPA; for Occitan, however, the most challenging target language in our experiments, MGPA beats all other approaches by a large margin: 17.12 with GPA versus 23.81 with MGPA. This pattern relates to the effect a supporting language has on the size of the induced seed dictionary. Figure 2 visualizes the progression of dictionary size during training with and without a supporting language for Occitan and Hebrew. The portion of the purple curves to the left of the dotted line corresponds to MGPA: notice how the curves are swapped between the two plots. Spanish *actually* provides support for the English-Occitan alignment, by contributing to an increasingly larger seed dictionary—this provides better anchoring for the learned alignment. Having Arabic as support for English-Hebrew alignment, on the other hand, causes a considerable reduction in the size of the seed dictionaries, giving GPA less anchor points and thus damaging the learned alignment. The variable effect of a supporting language on dictionary size, and consequently on alignment precision, relates to the quality of alignment of the support language with English and with the target language: referring back to Table 2, English-Spanish, for example, scores at 81.93, while English-Arabic precision is 35.33. Notice that despite our linguistically-motivated choice to pair related low- and high-resource languages for multi-support training, it is not necessarily the case that those should align especially well, as that would also depend on practical factors, such as embeddings quality and training corpora similarity (Søgaard et al., 2018).

MGPA+ applies two-way fine-tuning on top of MGPA. This leads to a drop in precision for Occitan, due to the removed support of Spanish and the consequent reduction in size of the induced dictionary (observe the fall of the purple curve after the dotted line in Figure 2 (a)). Meanwhile, precision for Hebrew is highest with MGPA+ out of all methods included. While Arabic itself is not a good support language, its presence in the three-way MGPA alignment seems to have resulted in a good initialization for the English-Hebrew two-way fine-tuning, thus helping the model reach an even better minimum along the loss curve.

5 Discussion: Why it works

If word vector spaces were completely isomorphic, the introduction of a third (or fourth) space, and the application of GPA, would lead to the same alignment as the alignment learned by PA, projecting the source language E into the target space F. This follows from the transitivity of isomorphism: if E is isomorphic to G and G is isomorphic to F, then E is isomorphic to F, via the isomorphism obtained by composing the isomorphisms from E to G and from G to F. So why do we observe improvements?

Søgaard et al. (2018) have shown that word vector spaces are often relatively far from being isomorphic, and approximate isomorphism is not transitive. What we observe therefore appears to be an instance of the Poincaré Paradox (Poincaré, 1902). While GPA is not more expressive than PA, it may still be easier to align each monolingual space to an intermediate space, as the latter constitutes a more similar target (albeit a non-isomorphic one); for example, the loss landscape of aligning a source and target language word embedding with an average of the two may be much smoother than when aligning source directly with target. Our work is in this way similar in spirit to Raiko et al. (2012), who use simple linear transforms to make learning of non-linear problems easier.

5.1 Error Analysis

Table 5 lists example translational pairs as induced from alignments between English and Bosnian, learned with PA, GPA and MGPA+. For interpretability, we query the system with words in Bosnian and seek their nearest neighbors in the English embedding space. P@1 over the Bosnian-English test set of Conneau et al. (2018) is 31.33, 34.80, and 34.47 for PA, GPA and MGPA+, respectively. The examples are grouped in three blocks, based on success and failure of PA and GPA alignments to retrieve a valid translation.

It appears that a lot of the difference in performance between PA and GPA concerns **morphologically related words**, e.g. *campaign* v. *campaigning*, *dialogue* v. *dialogues*, *merger* v. *merging* etc. These word pairs are naturally confusing to a BDI system, due to their related meaning and possibly identical syntactic properties (e.g. *merger* and *merging* can both be nouns). Another common mistake we observed in mismatches between PA and GPA predictions, was the wrong choice between two **antonyms**, e.g. *stable* v. *unstable* and *visible* v. *unnoticeable*. Distributional word representations are known to suffer from limitations with respect to capturing opposition of meaning (Mohammad et al., 2013), so it is not surprising that both PA- and GPA-learned alignments can fail in making this distinction. While it is not the case that GPA always outperforms PA on a query-to-query basis in these rather challenging cases, on average GPA appears to learn an alignment more robust to subtle morphological and semantic differences between neighboring words. Still, there are cases where PA and GPA both choose the wrong morphological variant of an otherwise correctly identified target word, e.g. *transformation* v. *transformations*.

Notice that many of the queries for which both algorithms fail, do result in a **nearly synonymous word** being predicted, e.g. *participant* for *attendee*, *earns* for *gets*, *footage* for *video*, etc. This serves to show that the learned alignments are generally good, but they are not sufficiently precise. This issue can have two sources: a suboptimal method for learning the alignment and/or a ceiling effect on how good of an alignment can be obtained, within the space of orthogonal linear transformations.

5.2 Procrustes fit

To explore the latter issue and to further compare the capabilities of PA and GPA, we perform a *Procrustes fit* test, where we learn alignments in a fully supervised fashion, using the test dictionaries of Conneau et al. (2018)[9] for both training *and* evaluation[10]. In the ideal case, i.e. if the subspaces defined by the words in the seed dictionaries are perfectly alignable, this setup should result in precision of 100%.

We found the difference between the fit with PA and GPA to be negligible, 0.20 on average across all 10 languages (5 low-resource and 5 high-source languages). It is not surprising that PA and GPA results in almost equivalent fits—the two algorithms both rely on linear transformations, i.e. they are equal in expressivity. As pointed out earlier, the superiority of GPA over PA stems from its

[9] For Occitan, we use our own test dictionary.

[10] In this experiment, we only run a single epoch of each alignment algorithm, as that is guaranteed to give us the best Procrustes fit for the particular set of training word pairs we would then evaluate on.

	QUERY	GOLD	PA	GPA	MGPA+
PA ✗, GPA ✓	variraju	vary	varies	vary	varies
	kanjon	canyon	headwaters	canyon	headwaters
	dijalog	dialogue	dialogues	dialogue	dialogue
	izjava	statement	deniable	statement	statements
	plazme	plasma	conduction	plasma	microspheres
	raunari	computers	minicomputers	computers	mainframes
	aparat	apparatus	duplex	apparatus	apparatus
	sazvijea	constellations	asterisms	constellations	constellations
	uspostavljanje	establishing	reestablishing	establishing	establishing
	industrijska	industrial	industry	industrial	industrial
	stabilna	stable	unstable	stable	stable
	disertaciju	dissertation	habilitation	dissertation	thesis
	protivnici	opponents	opposing	opponents	opponents
	pozitivni	positive	negative	positive	positive
	instalacija	installation	installations	installation	installation
	duhana	tobacco	liquors	tobacco	tobacco
PA ✓, GPA ✗	hor	choir	choir	musicum	choir
	crijevo	intestine	intestine	intestines	intestine
	vidljiva	visible	visible	unnoticeable	visible
	temelja	foundations	foundations	superstructures	pillars
	kolonijalne	colonial	colonial	colonialists	colonialists
	spajanje	merger	merger	merging	merging
	suha	dry	dry	humid	dry
	janez	janez	janez	mariza	janez
	kampanju	campaign	campaign	campaigning	campaign
	migracije	migration	migration	migrations	migrations
	sobu	room	room	bathroom	bathroom
	predgrau	suburb	suburb	outskirts	suburb
	specijalno	specially	specially	specialist	specially
	hiv	hiv	hiv	meningococcal	hiv
	otkrije	discover	discover	discovers	discover
	proizlazi	arises	arises	differentiates	deriving
	tajno	secretly	secretly	confidentially	secretly
PA ✗, GPA ✗	odred	squad	reconnoitre	stragglers	skirmished
	uesnik	attendee	participant	participant	participant
	saznao	learned	confided	confided	confided
	dobiva	gets	earns	earns	earns
	harris	harris	guinn	zachary	zachary
	snimke	videos	footage	footages	footage
	usne	lips	ear	ear	toes
	ukinuta	lifted	abolished	abolished	abolished
	objave	posts	publish	publish	publish
	obilježje	landmark	commemorates	commemorates	commemorates
	molim	please	appologize	thank	kindly
	vrste	solid	concretes	concretes	concretes
	intel	intel	genesys	motorola	transputer
	transformacije	transformations	transformation	transformation	transformation

Table 5: Example translations from Bosnian into English.

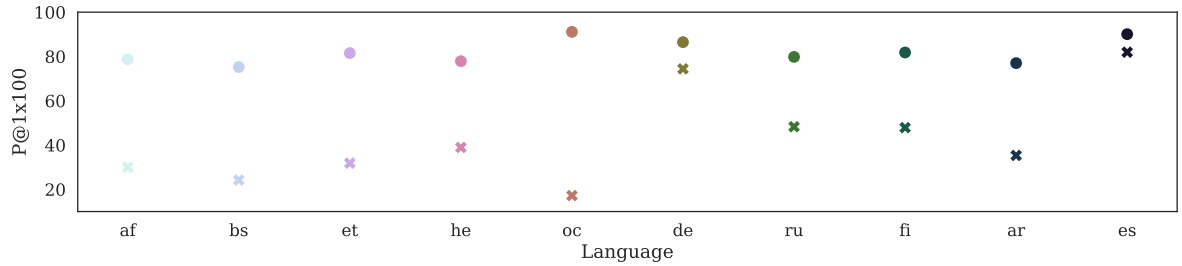


Figure 3: Procrustes fit test. Circles mark the results from fitting and evaluating GPA on the test dictionaries to measure the *Procrustes fit*. **x**s mark the weakly-supervised results reported in Tables 2 and 4.

more robust learning procedure, not from higher expressivity. Figure 3 thus only visualizes the Procrustes fit as obtained with GPA.

The Procrustes fit of all languages is indeed lower than 100%, showing that there is a **ceiling on the linear alignability** between the source and target spaces. We attribute this ceiling effect to variable degrees of linguistic difference between source and target language and possibly to differences in the contents of cross-lingual Wikipedias (recall that the embeddings we use are trained on Wikipedia corpora). An apparent correlation emerges between the Procrustes fit and precision scores for weakly-supervised GPA, i.e. between the circles and the **x**s in the plot. The only language that does not conform here is Occitan, which has the highest Procrustes fit and the lowest GPA precision out of all languages, but this result has an important caveat: our dictionary for Occitan comes from a different source and is much smaller than all the other dictionaries.

For some of the high-resource languages, weakly-supervised GPA takes us rather close to the best possible fit: e.g. for Spanish GPA scores 81.93%, and the Procrustes fit is 90.07%. While low-resource languages do not necessarily have lower Procrustes fits than high-resource ones (compare Estonian and Finnish, for example), the gap between the Procrustes fit and GPA precision is on average much higher within low-resource languages than within high-resource ones (52.46[11] compared to 25.47, respectively). This finding is in line with the common understanding that the quality of distributional word vectors depends on the amount of data available—we can infer from these results that suboptimal embeddings results in suboptimal cross-lingual alignments.

[11]Even if we leave Occitan out as an outlier, this number is still rather high: 47.10.

5.3 Multilinguality

Finally, we note that there may be specific advantages to including support languages for which large monolingual corpora exist, as those should, theoretically, be easier to align with English (also a high-resource language): variance in vector directionality, as studied in Mimno and Thompson (2017), increases with corpus size, so we would expect embedding spaces learned from corpora comparable in size, to also be more similar in shape.

6 Related work

Bilingual embeddings Many diverse cross-lingual word embedding models have been proposed (Ruder et al., 2018). The most popular kind learns a linear transformation from source to target language space (Mikolov et al., 2013). In most recent work, this mapping is constrained to be orthogonal and solved using Procrustes Analysis (Xing et al., 2015; Artetxe et al., 2017, 2018; Conneau et al., 2018; Lu et al., 2015). The approach most similar to ours, Faruqui and Dyer (2014), uses canonical correlation analysis (CCA) to project both source and target language spaces into a third, joint space. In this setup, similarly to GPA, the third space is iteratively updated, such that at timestep t, it is a product of the two language spaces as transformed by the mapping learned at timestep $t-1$. The objective that drives the updates of the mapping matrices is to maximize the correlation between the projected embeddings of translational equivalents (where the latter are taken from a gold-standard seed dictionary). In their analysis of the transformed embedding spaces, Faruqui and Dyer (2014) focus on the improved quality of monolingual embedding spaces themselves and do not perform evaluation

of the task of BDI. They find that the transformed monolingual spaces better encode the difference between synonyms and antonyms: in the original monolingual English space, synonyms and antonyms of *beautiful* are all mapped close to each other in a mixed fashion; in the transformed space the synonyms of *beautiful* are mapped in a cluster around the query word and its antonyms are mapped in a separate cluster. This finding is in line with our observation that GPA-learned alignments are more precise in distinguishing between synonyms and antonyms.

Multilingual embeddings Several approaches extend existing methods to space alignments between more than two languages (Ammar et al., 2016; Ruder et al., 2018). Smith et al. (2017) project all vocabularies into the English space. In some cases, multilingual training has been shown to lead to improvements over bilingually trained embedding spaces (Vulić et al., 2017), similar to our findings.

7 Conclusion

Generalized Procrustes Analysis yields benefits over simple Procrustes Analysis for Bilingual Dictionary Induction, due to its smoother loss landscape. In line with earlier research, benefits from the introduction of a common latent space seem to relate to a better distinction of synonyms and antonyms, and of syntactically-related words. GPA also offers the possibility to include multilingual support for inducing a larger seed dictionary during training, which better anchors the English to target language alignment in low-resource scenarios.

Acknowledgements

Sebastian is supported by Irish Research Council Grant Number EBPPG/2014/30 and Science Foundation Ireland Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively Multilingual Word Embeddings.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations. In *Proceedings of AAAI 2018*.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *Proceedings of ICLR 2018*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving Zero-Shot Learning by Mitigating the Hubness Problem. *ICLR 2015 Workshop track*, pages 1–10.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. *arXiv preprint arXiv:1808.08780*.

Manaal Faruqui and Chris Dyer. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462 – 471.

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

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep Multilingual Correlation for Improved Word Embeddings. In *HLT-NAACL*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of EMNLP*.

Saif M. Mohammad, Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.

Henri Poincaré. 1902. *La Science et 1' Hypothese*. Flammario, Paris, France.

Tapani Raiko, Harri Valpola, and Yann LeCun. 2012. Deep learning made easier by linear transformations in perceptrons. In *AISTATS*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2018. A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*.

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

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of ACL 2018*.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017. Cross-Lingual Induction and Transfer of Verb Classes Based on Word Vector Space Specialisation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Chao Xing, Chao Liu, Dong Wang, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. *NAACL-2015*, pages 1005–1010.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 1999. On the universal structure of human lexical semantics. In *NIPS*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial Training for Unsupervised Bilingual Lexicon Induction. In *Proceedings of ACL*.