Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations

Mikel Artetxe and **Gorka Labaka** and **Eneko Agirre**
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Abstract

Using a dictionary to map independently trained word embeddings to a shared space has shown to be an effective approach to learn bilingual word embeddings. In this work, we propose a multi-step framework of linear transformations that generalizes a substantial body of previous work. The core step of the framework is an orthogonal transformation, and existing methods can be explained in terms of the additional normalization, whitening, re-weighting, de-whitening and dimensionality reduction steps. This allows us to gain new insights into the behavior of existing methods, including the effectiveness of inverse regression, and design a novel variant that obtains the best published results in zero-shot bilingual lexicon extraction. The corresponding software is released as an open source project.

1 Introduction

Bilingual word embeddings have attracted a lot of attention in recent times. Most methods to learn them use some sort of bilingual signal at the document level, either in the form of document-aligned or label-aligned comparable corpora (Søgaard et al. 2015; Vulić and Moens 2016; Mogadala and Rettinger 2016) or, more commonly, in the form of parallel corpora (Gouws, Bengio, and Corrado 2015; Luong, Pham, and Manning 2015).

An alternative approach that we address in this paper is to independently train the embeddings for each language on monolingual corpora, and then map them to a shared space based on a bilingual dictionary (Mikolov, Le, and Sutskever 2013; Lazaridou, Dinu, and Baroni 2015). This requires minimal bilingual supervision compared to other approaches, while allowing to leverage large amounts of monolingual corpora with competitive results (Vulić and Korhonen 2016; Artetxe, Labaka, and Agirre 2017). Moreover, the learned mappings can also be applied to words that were missing in the training dictionary, and thus induce their translations, with improvements in machine translation (Zhao, Hassan, and Auli 2015).

Authors have proposed different methods to learn such word embedding mappings, but their approach and motivations are often divergent, making it difficult to get a general understanding of the topic. In this work, we tackle this issue and propose a multi-step framework that generalizes previous work. The core step of the framework, which maps both languages to a shared space using an orthogonal transformation, is shared by all variants, and the differences between previous methods are exclusively explained in terms of their normalization, whitening, re-weighting and dimensionality reduction behavior. We analyze the effect of each of these steps with experimental support, which allows us to gain new insights into the behavior of existing methods. Based on these insights, we design a novel variant that improves the state-of-the-art in bilingual lexicon extraction.

Our framework is highly related to the zero-shot learning paradigm, where a multi-class classifier trained over a subset of the labels learns to predict unseen labels by exploiting a common representation for them (Palatucci et al. 2009). In our scenario, these labels correspond to the target language words and their common representation is provided by their corresponding embeddings. This is a prototypical zero-shot learning problem, and similar mapping techniques have also been used in other zero-shot tasks like image labeling (Shigeto et al. 2015; Lazaridou, Dinu, and Baroni 2015) and drug discovery (Larochelle, Erhan, and Bengio 2008).

The remaining of this paper is organized as follows. Section 2 discusses related work. Section 3 explains the proposed multi-step framework and shows the equivalence with previous methods. Section 4 then presents the experimental settings, while Section 5 discusses the obtained results. Section 6 concludes the paper.

2 Related work

For the sake of space, we will focus on related work directly relevant to embedding mappings and bilingual lexicon extraction. Bilingual embedding mapping methods work by independently training the word embeddings in two languages, and then mapping them to a shared space based on a bilingual dictionary. Even if the literature in the topic is quite broad, existing methods can be classified in the following four groups:

1. **Regression methods** map the embeddings in one language to maximize their similarity with the other language. For that purpose, methods in this group use a least-squares objective function that learns the linear transformation minimizing the sum of squared Euclidean distances for the dictionary entries. This approach was first

proposed by Mikolov, Le, and Sutskever (2013), and later adopted by many other authors that incorporated L2 regularization (Dinu, Lazaridou, and Baroni 2015; Lazaridou, Dinu, and Baroni 2015; Vulić and Korhonen 2016). Even if the linear transformation is usually learned from the source language into the target language, Shigeto et al. (2015) argue that it is better to map the target language into the source language as a way to address the hubness problem[1].

2. **Canonical methods** map the embeddings in both languages to a shared space where their similarity is maximized. This is usually done through Canonical Correlation Analysis (CCA) as first proposed by Faruqui and Dyer (2014), who motivate their method as a way to improve the quality of monolingual embeddings using bilingual data. With a similar motivation, Lu et al. (2015) extend this work and use Deep Canonical Correlation Analysis to learn non-linear mappings. CCA was also extended to the multilingual scenario by Ammar et al. (2016) taking English as the pivot language.

3. **Orthogonal methods** map the embeddings in one or both languages to maximize their similarity, but constrain the transformation to be orthogonal. This constraint has been introduced with different motivations. Xing et al. (2015) allege inconsistencies in previous approaches, and orthogonality serves to preserve the length normalization performed by their method to address them. Artetxe, Labaka, and Agirre (2016) motivate orthogonality as a way to preserve monolingual invariance, preventing the degradation in monolingual tasks observed for other techniques. Zhang et al. (2016) focus on a transfer-learning scenario with only ten translation pairs for training, and incorporate orthogonality as a hard regularizer. Finally, Smith et al. (2017) point out that the mapping should be orthogonal in order to be self-consistent.

4. **Margin methods** map the embeddings in one language to maximize the margin between the correct translations and the rest of the candidates. This approach was proposed by Lazaridou, Dinu, and Baroni (2015) as a way to address the hubness problem, with the addition of intruder negative sampling to generate more informative training examples.

As it can be seen, the previous work on embedding mappings is quite diverse, with many authors working under different scenarios and motivations. In an attempt to provide a more general view, Artetxe, Labaka, and Agirre (2016) show the equivalence of different objective functions under orthogonality and different normalization procedures, and clarify that regression, canonical and orthogonal methods essentially differ on the constraints imposed on the mapping. In contrast, our framework decomposes these differences into several interpretable steps, which allows us to gain additional insights into the behavior of previous methods and design new variants addressing their deficiencies. We also cover additional methods, including most references in this section (see Table 1).

A practical application of embedding mappings, as well as the main evaluation task, is bilingual lexicon extraction, that is, the zero-shot translation of words that were missing in the training dictionary. This is usually done through **nearest neighbor retrieval**, taking the closest embedding in the target language according to some similarity metric (usually cosine). However, Dinu, Lazaridou, and Baroni (2015) argue that this approach suffers from the hubness problem, and propose using **inverted nearest neighbor retrieval**[2] instead, which takes the target embedding that has the source embedding ranked highest in its nearest neighbor list. Ties are solved by taking the candidate with the highest cosine similarity. Finally, **inverted softmax retrieval** (Smith et al. 2017) also works by reversing the direction of the query, but instead of using the cosine in the similarity computations, it uses a softmax function with a hyperparameter to control the temperature, which is tuned in the training dictionary. In this paper we revisit these techniques, and show that the alternatives to nearest neighbor mitigated deficiencies in previous mapping methods, while our method learns better mappings.

3 Proposed framework

Let X and Z be the word embedding matrices in two languages for a given bilingual dictionary so that their ith row X_{i*} and Z_{i*} are the embeddings of the ith entry. We aim to learn the transformation matrices W_X and W_Z so the mapped embeddings XW_X and ZW_Z are close to each other.

We next propose a multi-step framework to learn such mappings that allows to generalize previous work. The ith step of the framework applies a linear transformation to the output embeddings of the previous step in each language. This way, if $X_{(i)}$ denotes the output embeddings in the source language at step i and $W_{X(i)}$ the linear transformation at step i, we will have $X_{(i)} = X_{(i-1)}W_{X(i)}$ and $W_X = \prod_i W_{X(i)}$, and analogously for the target language. As it is clear from this last expression, the composition of several linear transformations is another linear transformation, so the purpose of our framework is not to improve the expressive power of linear mappings, but rather to decompose them into several meaningful steps. More concretely, our framework consists of the following steps:

- **Step 0: Normalization (optional)**: In this optional preprocessing step, the word embeddings in each language are independently normalized. This can involve length normalization (making all embeddings have a unit Euclidean norm), and mean centering (making each component have a zero mean). Note that this is done as a preprocessing step, obtaining the initial embedding matrices $X_{(0)}$ and $Z_{(0)}$ that will be mapped by the following ones.

[1] Hubness (Radovanović, Nanopoulos, and Ivanović 2010a; 2010b) refers to the phenomenon of some points (known as *hubs*) being the nearest neighbors of many other points in high-dimensional spaces, and has been reported to severely affect bilingual embedding mappings (Dinu, Lazaridou, and Baroni 2015; Lazaridou, Dinu, and Baroni 2015; Shigeto et al. 2015; Smith et al. 2017).

[2] Note that the original paper refers to this method as *globally-corrected* retrieval.

		S0 (l)	S0 (m)	S1	S2	S3	S4 (src)	S4 (trg)	S5
OLS	Mikolov, Le, and Sutskever (2013)			x	x	src	trg	trg	
	Shigeto et al. (2015)			x	x	trg	src	src	
CCA	Faruqui and Dyer (2014)	x	x	x	x				x
Orth.	Xing et al. (2015)	x			x				
	Zhang et al. (2016)				x				
	Artetxe, Labaka, and Agirre (2016)	x	x		x				
	Smith et al. (2017)	x			x				x
	Proposed (Section 5)	x	x	x	x	trg	src	trg	x

Table 1: Equivalence of the proposed framework with previous methods. (l) and (m) denote length normalization and mean centering, respectively.

- **Step 1: Whitening (optional)**. This optional step applies a whitening or sphering transformation to the embeddings in each language, which makes their different components have a unit variance and be uncorrelated among themselves, turning their covariance matrices into the identity matrix[3]. For that purpose, we adopt the Mahalanobis or ZCA whitening, taking $W_{X(1)} = (X^T X)^{-\frac{1}{2}}$ and $W_{Z(1)} = (Z^T Z)^{-\frac{1}{2}}$.
- **Step 2: Orthogonal mapping**. This step maps the embeddings in both languages to a shared space. Both transformations are constrained to be orthogonal, preserving the dot product for each of the languages on their own. More concretely, we take $W_{X(2)} = U$ and $W_{Z(2)} = V$, where $USV^T = X_{(1)}^T Z_{(1)}$ is the SVD factorization of $X_{(1)}^T Z_{(1)}$. This maximizes the summative cross-covariance of the mapped embeddings $\mathrm{Tr}(X_{(1)} W_{X(2)} W_{Z(2)}^T Z_{(1)}^T)$. Moreover, the ith component of the mapped embeddings corresponds to the direction of maximum cross-covariance being orthogonal to the previous ones, and S_{ii} is its corresponding cross-covariance value. Note that when whitening is applied at step 1, the variance in all directions is 1, so the cross-covariance is equivalent to the cross-correlation.
- **Step 3: Re-weighting (optional)**: This optional step re-weights each component according to its cross-correlation, increasing the relevance of those that best match across languages. So as to simplify the formalization, we will only consider this step if step 1 was applied before, in which case the cross-correlations correspond to the singular values in S (step 2). The re-weighting can be applied to the source language embeddings ($W_{X(3)} = S$ and $W_{Z(3)} = I$), or to the target language embeddings ($W_{X(3)} = I$ and $W_{Z(3)} = S$).
- **Step 4: De-whitening (optional)**: This optional step restores the original variance in every direction, and it is thus only meaningful if step 1 was applied before. The embeddings in a given language can be de-whitened with respect to the original variance in that same language, but also with respect to the original variance in the other language, as both languages are in the same space after step 2. In either case, de-whitening language A with respect to B requires $W_{A(4)} = W_{B(2)}^T W_{B(1)}^{-1} W_{B(2)}$.
- **Step 5: Dimensionality reduction (optional)**: This optional step keeps the first n components of the resulting embeddings and drops the rest, which is obtained by $W_{X(5)} = W_{Z(5)} = (I_n \ 0)^T$. This can be seen as an extreme form of re-weighting, where the first n components are re-weighted by one and the remaining ones by zero.

An interesting aspect of this framework is that the mapping of both languages to a common space is reduced to a single step that is shared by all variants (step 2). Moreover, this mapping is orthogonal and, therefore, preserves monolingual invariance. Therefore, different variants, including existing methods, will only differ on their treatment of normalization, whitening/de-whitening, re-weighting and dimensionality reduction, which are easier to interpret. More concretely, the equivalence of this framework with existing methods, detailed in Table 1, is as follows:

- **Regression methods** correspond to the case where both languages are whitened, re-weighting is applied to the source language, and both languages are de-whitened with respect to the target language (or inversely if the regression is applied from the target language into the source language). This equivalence is directly given by the close-form solution of the unregularized variant, known as Ordinary Least Squares (OLS)[4], and we leave the analysis of $L2$ regularization for future work.

[3]Note that our use of the variance and covariance concepts at this step and the following ones assumes that the embeddings are already mean centered (i.e. we take $X^T X$ as (proportional to) the covariance matrix of X, $Z^T Z$ as the covariance matrix of Z, and $X^T Z$ as the cross-covariance matrix of X and Z).

[4]The optimal solution of OLS is given by $W_{OLS} = X^+ Z$, where $X^+ = \left(X^T X\right)^{-1} X^T$ is the Moore-Penrose pseudoinverse of X. At the same time, by simple algebraic development of our claimed equivalence, $W_X = \left(X^T X\right)^{-1} X^T Z V$ and $W_Z = V$, where V is an orthogonal matrix given by the SVD factorization at step 2. Therefore, both solutions are equivalent up to the orthogonal transformation V of the resulting space, which is invariant with respect to the dot product.

- **Canonical methods** (CCA) correspond to the case where both languages are whitened, none is de-whitened, re-weighting is not used, and dimensionality reduction is applied. The equivalence is given by the SVD solution of CCA (see for instance Lu and Foster (2014)).
- **Orthogonal methods** correspond to the simplest case without any whitening, re-weighting and de-whitening. The equivalence is directly given by the transformation learned at step 2, which is equivalent to the solutions of Artetxe, Labaka, and Agirre (2016) and Smith et al. (2017).

As it can be seen, our framework covers all mapping families with the exception of margin based ones, which were only explored by Lazaridou, Dinu, and Baroni (2015) and surpassed by subsequent work.

4 Experimental settings

For easier comparison with related work, we performed our experiments in the bilingual lexicon extraction scenario proposed by Dinu, Lazaridou, and Baroni (2015) and used by subsequent authors. Their public English-Italian dataset[5] includes monolingual word embeddings in both languages together with a bilingual dictionary split in a training set and a test set. Artetxe, Labaka, and Agirre (2017) extended this dataset to English-German and English-Finnish, which we also use in our experiments. In all cases, the embeddings were trained with the word2vec toolkit with CBOW and negative sampling (Mikolov et al. 2013)[6]. The training and test sets were derived from dictionaries built from Europarl word alignments and available at OPUS (Tiedemann 2012), taking 1,500 random entries uniformly distributed in 5 frequency bins as the test set and the 5,000 most frequent pairs of the remaining word pairs as the training set. The corpora used consisted of 2.8 billion words for English (ukWaC + Wikipedia + BNC), 1.6 billion words for Italian (itWaC), 0.9 billion words for German (SdeWaC), and 2.8 billion words for Finnish (Common Crawl from WMT 2016).

In addition to these languages, we further extended the dataset to English-Spanish using the exact same settings described above. For that purpose, we used the WMT News Crawl 2007-2012 corpus[7] for Spanish, which consists of 386 million words. Tokenization was performed using standard Moses tools. Note that the resulting Spanish corpus has a different domain to the previous ones (news vs web crawling), and it is also smaller, which explains the lower accuracy numbers in the next section.

The goal of our experiments is twofold. On the one hand, we want to analyze the effect of each of the steps of our framework on their own, and interpret the results in relation to the behavior of previous methods. On the other hand, we want to identify the best variant of our framework, and compare it with existing methods proposed in the literature. Given that the effect of normalization was already analyzed in detail by Artetxe, Labaka, and Agirre (2016), we leave this factor aside in our experiments and use their recommended configuration, which performs length normalization followed by mean centering. Moreover, we use cosine similarity with standard nearest neighbor as our retrieval method unless otherwise specified, which allows us to better evaluate the quality of the mapping itself. The remaining factors are analyzed independently, and their best combination is then compared to the state-of-the-art. The code and resources to reproduce our experiments are available at `https://github.com/artetxem/vecmap`.

[5] `http://clic.cimec.unitn.it/~georgiana.dinu/down/`

[6] The context window was set to 5 words, the dimension of the embeddings to 300, the sub-sampling to 1e-05 and the number of negative samples to 10, and the vocabulary was restricted to the 200,000 most frequent words.

[7] `http://www.statmt.org/wmt13/translation-task.html`

5 Results and discussion

From Section 5.1 to 5.4, we respectively analyze the effect of whitening/de-whitening, re-weighting, dimensionality reduction and the retrieval method. Section 5.5 then compares the proposed system with other methods in the literature.

5.1 Whitening and de-whitening (steps 1 and 4)

As discussed before, existing methods have a very different behavior with respect to whitening. While orthogonal methods do not perform any whitening, both CCA and OLS whiten both languages, and the latter also de-whitens them with respect to one of the languages, depending on the direction of regression (see Table 1).

Table 2 shows our results for different whitening/de-whitening strategies. In addition to the said variants implicitly used by existing methods, it also includes our proposed variant: the more intuitive choice of de-whitening each language with respect to the original variance in that same language.

As it can be seen, the results show that, for most language pairs, whitening and de-whitening each language with respect to itself brings a small improvement over not whitening at all. The only exception is English-Finnish, whose accuracy drops almost one point with respect to not applying any whitening or de-whitening. A possible explanation of why proper whitening and de-whitening helps is a hypothetical bias that would otherwise push directions with high variance together.

But, more importantly, the results show that the whitening/de-whitening behavior of both CCA and OLS is not only counterintuitive, but also harmful. In the case of the former, simply whitening both languages causes a huge accuracy drop of 7-9 points, suggesting that the variances of the original embeddings are relevant and should not be ignored by any means. In the case of the latter, de-whitening with respect to either language causes an accuracy drop of 2-4 points, showing that this de-whitening strategy is better than not de-whitening at all, but worse than the natural choice of de-whitening with respect to the language in question.

5.2 Re-weighting (step 3)

As seen in the Section 3, neither orthogonal methods nor CCA use re-weighting, while OLS re-weights either the

Motivation	S1	S4 (src)	S4 (trg)	EN-IT	EN-DE	EN-FI	EN-ES
Orth.				39.27%	41.87%	**30.62%**	31.40%
CCA	x			32.27%	33.00%	22.05%	23.73%
OLS	x	src	src	37.33%	38.47%	25.35%	28.87%
	x	trg	trg	38.00%	36.60%	26.33%	28.80%
New	x	src	trg	**39.47%**	**41.93%**	29.71%	**31.67%**

Table 2: Accuracy for different whitening (S1) and de-whitening (S4) configurations. All settings use length normalization and mean centering, and do not re-weight nor apply dimensionality reduction.

Mot.	S3	EN-IT	EN-DE	EN-FI	EN-ES
Orth. / CCA		39.47%	41.93%	29.71%	31.67%
OLS	src	38.53%	41.73%	28.65%	30.47%
	trg	**43.80%**	**44.27%**	**32.79%**	**36.47%**

Table 3: Accuracy for different re-weighting (S3) configurations. All settings use length normalization, mean centering, and whitening/de-whitening with respect to the original language.

S3	S5	EN-IT	EN-DE	EN-FI	EN-ES
		39.47%	41.93%	29.71%	31.67%
	x	42.53%	**44.53%**	32.09%	33.80%
trg		43.80%	44.27%	32.79%	36.47%
	x	**44.00%**	44.27%	**32.94%**	**36.53%**

Table 4: Accuracy for different dimensionality reduction (S5) and re-weighting (S3) configurations. All settings use length normalization, mean centering, whitening, and de-whitening with respect to the original language.

source or the target language depending on the direction of regression. Table 3 shows the results obtained for all these different re-weighting strategies.

As it can be seen, re-weighting the target language is highly beneficial, bringing an improvement of 3-5 points in all cases, while re-weighting the source language is always harmful. Interestingly, which side to re-weight should not be a relevant factor when using the dot product, so this difference must be explained by the length normalization performed by cosine similarity. Note that, when re-weighting the source language, this length normalization is applied to each source language word on its own, but its nearest neighbor list is not affected in any way, as its similarity with respect to all target language words is only scaled by a constant normalization factor. As a consequence, for the length normalization of cosine similarity to be effective in nearest neighbor retrieval, the re-weighting must be applied in the target language, which can explain why we obtain better results for it.

This behavior is also consistent with the findings of Shigeto et al. (2015) regarding the direction of regression. Recall that these authors claim that mapping the target language into the source language is better than mapping the source language into the target language, which respectively correspond to re-weighting the target language and the source language according to our framework (see Table 1). While Shigeto et al. (2015) explain the relevance of the regression direction in terms of the emergence of hubs in the subsequent nearest neighbor retrieval, our work identifies that the origin of this problem is in the implicit re-weighting direction and its relation with the length normalization performed by cosine similarity.

5.3 Dimensionality reduction (step 5)

As discussed before, CCA is always used with dimensionality reduction, while OLS never is. Dimensionality reduction is typically not applied in orthogonal methods either, although Smith et al. (2017) recently introduced it for the first time.

Table 4 shows our results with and without dimensionality reduction. When performing dimensionality reduction, we always chose the number of dimensions that yield the highest accuracy in the training dictionary, and then evaluate in the test set. As discussed in Section 3, dimensionality reduction can be seen as an extreme form of re-weighting, so we performed these experiments with and without re-weighting the target language so as to better understand how these two steps interact.

As it can be seen, dimensionality reduction has a positive effect in all cases. However, its impact is very small when using target language re-weighting (an improvement of 0.20 points in the best case), and much bigger when not using any re-weighting (improvements of 2-3 points). This suggests that re-weighting and dimensionality reduction have an overlapping effect, which reinforces our interpretation that dimensionality reduction is just an extreme form of re-weighting that removes the components with smallest cross-correlation. In relation to that, it is remarkable that re-weighting gives considerably better results than dimensionality reduction alone, which can be attributed to its smooth rescaling of embedding components in contrast to the binary discarding performed by dimensionality reduction. The only exception in this regard is English-German, for which dimensionality reduction alone gives slightly better results.

All in all, we can conclude that, in spite of being con-

Retrieval method	EN-IT	EN-DE	EN-FI	EN-ES
Nearest neighbor	44.00%	**44.27%**	**32.94%**	36.53%
Inverted nearest neighbor	43.07%	42.20%	31.18%	32.53%
Inverted softmax	**45.27%**	44.13%	**32.94%**	**36.60%**

Table 5: Accuracy for different retrieval methods. All settings use length normalization, mean centering, whitening, target language re-weighting, de-whitening with respect to the original language, and dimensionality reduction tuned in training.

	EN-IT	EN-DE	EN-FI	EN-ES
Mikolov, Le, and Sutskever (2013)	34.93% (**)	35.00% (**)	25.91% (**)	27.73% (**)
Faruqui and Dyer (2014)	38.40% (*)	37.13% (*)	27.60% (*)	26.80% (*)
Shigeto et al. (2015)	41.53% (**)	43.07% (**)	31.04% (**)	33.73% (**)
Dinu, Lazaridou, and Baroni (2015)	37.7% / 38.53% (*)	38.93% (*)	29.14% (*)	30.40% (*)
Lazaridou, Dinu, and Baroni (2015)	40.2%	-	-	-
Xing et al. (2015)	36.87% (**)	41.27% (**)	28.23% (**)	31.20% (**)
Zhang et al. (2016)	36.73% (**)	40.80% (**)	28.16% (**)	31.07% (**)
Artetxe, Labaka, and Agirre (2016)	39.27%	41.87% (*)	30.62% (*)	31.40% (*)
Smith et al. (2017)	43.1% / 44.53% (**)	43.33% (**)	29.42% (**)	35.13% (**)
Proposed (nearest neighbor)	44.00%	**44.27%**	**32.94%**	36.53%
Proposed (inverted softmax)	**45.27%**	44.13%	**32.94%**	**36.60%**

Table 6: Accuracy of our method in comparison with previous work. (*) means that the results were obtained using the original implementation from the authors, while (**) means that the results were obtained using our custom implementation as part of our proposed framework. The rest of the results were reported in the original papers. For methods that were not originally proposed for bilingual lexicon extraction, we used nearest neighbor retrieval.

nected, re-weighting tends to work considerably better than dimensionality reduction thanks to its smooth nature. Moreover, combining them has a small but positive impact, and should be the preferred configuration to use.

5.4 Retrieval method

Most previous work uses standard nearest neighbor for bilingual lexicon extraction (see Section 2), but alternative retrieval methods have been proposed to address the hubness problem attributed to it (Dinu, Lazaridou, and Baroni 2015; Smith et al. 2017). Table 5 reports the results for each of these methods. In the case of inverted softmax, we tune the inverse temperature to optimize the accuracy in the training set, which we find to work better than maximizing the log-likelihood as originally proposed by Smith et al. (2017). To speed up the computations, we take a random sample of 1,500 words to estimate the partition function of the softmax during tuning, but use the entire source vocabulary in the test set. Similarly, we use the entire source vocabulary as pivots when using inverted nearest neighbor.

As it can be seen, inverted softmax performs at par with standard nearest neighbor retrieval for all language pairs except for English-Italian, where it brings an improvement of 1.27 points. Note that this number is considerably smaller than the nearly 5 points reported by Smith et al. (2017) for the same dataset. At the same time, inverted nearest neighbor performs worse than standard nearest neighbor in our experiments. This suggests that alternative retrieval methods are not fully complementary with the improvements brought by our framework. We hypothesize that this is connected to our previous discussion on re-weighting. Recall that our work explains that, for the length normalization performed by cosine similarity to be effective in nearest neighbor retrieval, the re-weighting should be performed in the opposite side. Nevertheless, most previous work was not applying re-weighting properly, and alternative retrieval methods would mitigate the problem by reversing the direction of nearest neighbor. Note, thus, that the alternative methods were alleviating an inherent flaw of the mapping methods during retrieval, while our framework learns better mappings.

5.5 Comparison with the state-of-the-art

Having analyzed the different steps of the proposed framework on their own, we next analyze how it performs in comparison to other methods proposed in the literature. For that purpose, we choose the recommended variant of our framework as discussed throughout the section, which is the one using whitening, re-weighting the target language, de-whitening with respect the original language, and applying dimensionality reduction (see Table 1). The obtained results are given in Table 6. Note that we only tried limited combinations of well-motivated steps and, given that we tested in several pairs of languages, we think that our conclusions are well supported. Moreover, note that our implementation of inverted softmax optimizes accuracy and uses the entire source vocabulary for computing the partition function at test time as described in Section 5.4, which performs better than the variant reported in Smith et al. (2017) as shown

by its corresponding line in the table (43.1 vs. 44.53 for EN-IT).

As it can be seen, our system obtains the best published results in all the four language pairs. Moreover, it also surpasses the previous state-of-the-art even when using standard nearest neighbor retrieval, which shows the superiority of the mapping method itself.

6 Conclusions and future work

In this work, we propose a new framework to learn bilingual embedding mappings that generalizes a substantial body of previous work (Mikolov, Le, and Sutskever 2013; Faruqui and Dyer 2014; Shigeto et al. 2015; Xing et al. 2015; Zhang et al. 2016; Artetxe, Labaka, and Agirre 2016; Smith et al. 2017). A key aspect of our framework is that the mapping to a common space is reduced to a single orthogonal transformation that is shared by all variants, and their differences are exclusively explained in terms of their normalization, whitening, re-weighting, de-whitening and dimensionality reduction behavior. This allows us to gain new insights into existing mapping methods, as follows:

- Whitening can bring small improvements, but only if de-whitened appropriately. Our work shows that the implicit de-whitening behavior of both OLS methods (Mikolov, Le, and Sutskever 2013) and CCA methods (Faruqui and Dyer 2014) is flawed.
- Re-weighting is very helpful, but, contrary to most previous work, it should be performed in the target language for the length normalization performed by cosine similarity to be effective in nearest neighbor retrieval. This explains why mapping the target language into the source language performs better than mapping the source language into the target language for regression methods (Shigeto et al. 2015).
- Dimensionality reduction is an extreme form of re-weighting. Even if it was shown to be beneficial with CCA methods (Faruqui and Dyer 2014) and orthogonal methods (Smith et al. 2017), smooth re-weighting gives even better results. Using both of them together is not harmful, bringing further improvements in some cases, and should be the default configuration to try.

Moreover, we also shed light on the relation between mapping methods and retrieval methods when inducing bilingual lexicons:

- The use of alternative retrieval methods to nearest neighbor (Dinu, Lazaridou, and Baroni 2015; Smith et al. 2017) mitigated deficiencies in the implicit re-weighting behavior of previous mapping methods. When re-weighting is properly applied in the target language, inverted softmax (Smith et al. 2017) performs at par with standard nearest neighbor in most cases, while inverted nearest neighbor gives considerably worse results.

Based on these insights, we propose a new variant that obtains the best published results in bilingual lexicon extraction for all the four language pairs tested. We release our implementation as an open source project, which allows to replicate several previous methods as well as our improved variant (Mikolov, Le, and Sutskever 2013; Faruqui and Dyer 2014; Dinu, Lazaridou, and Baroni 2015; Shigeto et al. 2015; Xing et al. 2015; Zhang et al. 2016; Smith et al. 2017; Artetxe, Labaka, and Agirre 2016). In the future, we would like to incorporate L2 regularization in our framework and extend our analysis to max-margin methods (Lazaridou, Dinu, and Baroni 2015) and non-linear mappings (Lu et al. 2015). Moreover, we would like to introduce hyperparameters to control the intensity of whitening/de-whitening and re-weighting, which we believe could bring further improvements with proper tuning. Finally, we would like to adapt and evaluate our framework in other zero-shot learning scenarios.

Acknowledgments

This research was partially supported by a Google Faculty Award, the Spanish MINECO (TUNER TIN2015-65308-C5-1-R, MUSTER PCIN-2015-226 and TADEEP TIN2015-70214-P, cofunded by EU FEDER), the Basque Government (MODELA KK-2016/00082) and the UPV/EHU (excellence research group). Mikel Artetxe enjoys a doctoral grant from the Spanish MECD.

References

Ammar, W.; Mulcaire, G.; Tsvetkov, Y.; Lample, G.; Dyer, C.; and Smith, N. A. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Artetxe, M.; Labaka, G.; and Agirre, E. 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*, 2289–2294. Austin, Texas: Association for Computational Linguistics.

Artetxe, M.; Labaka, G.; and Agirre, E. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 451–462. Vancouver, Canada: Association for Computational Linguistics.

Dinu, G.; Lazaridou, A.; and Baroni, M. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), workshop track*.

Faruqui, M., and Dyer, C. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 462–471. Gothenburg, Sweden: Association for Computational Linguistics.

Gouws, S.; Bengio, Y.; and Corrado, G. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, 748–756.

Larochelle, H.; Erhan, D.; and Bengio, Y. 2008. Zero-data learning of new tasks. In *AAAI Conference on Artificial Intelligence*.

Lazaridou, A.; Dinu, G.; and Baroni, M. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 270–280. Beijing, China: Association for Computational Linguistics.

Lu, Y., and Foster, D. P. 2014. Large scale canonical correlation analysis with iterative least squares. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 91–99.

Lu, A.; Wang, W.; Bansal, M.; Gimpel, K.; and Livescu, K. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 250–256. Denver, Colorado: Association for Computational Linguistics.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 151–159. Denver, Colorado: Association for Computational Linguistics.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc. 3111–3119.

Mikolov, T.; Le, Q. V.; and Sutskever, I. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mogadala, A., and Rettinger, A. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 692–702. San Diego, California: Association for Computational Linguistics.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In Bengio, Y.; Schuurmans, D.; Lafferty, J. D.; Williams, C. K. I.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc. 1410–1418.

Radovanović, M.; Nanopoulos, A.; and Ivanović, M. 2010a. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research* 11(Sep):2487–2531.

Radovanović, M.; Nanopoulos, A.; and Ivanović, M. 2010b. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 186–193. ACM.

Shigeto, Y.; Suzuki, I.; Hara, K.; Shimbo, M.; and Matsumoto, Y. 2015. *Ridge Regression, Hubness, and Zero-Shot Learning*. Cham: Springer International Publishing. 135–151.

Smith, S. L.; Turban, D. H.; Hamblin, S.; and Hammerla, N. Y. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations (ICLR 2017)*.

Søgaard, A.; Agić, v.; Martínez Alonso, H.; Plank, B.; Bohnet, B.; and Johannsen, A. 2015. Inverted indexing for cross-lingual NLP. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1713–1722. Beijing, China: Association for Computational Linguistics.

Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Vulić, I., and Korhonen, A. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 247–257. Berlin, Germany: Association for Computational Linguistics.

Vulić, I., and Moens, M.-F. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research* 55(1):953–994.

Xing, C.; Wang, D.; Liu, C.; and Lin, Y. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1006–1011. Denver, Colorado: Association for Computational Linguistics.

Zhang, Y.; Gaddy, D.; Barzilay, R.; and Jaakkola, T. 2016. Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1307–1317. San Diego, California: Association for Computational Linguistics.

Zhao, K.; Hassan, H.; and Auli, M. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1527–1536. Denver, Colorado: Association for Computational Linguistics.