A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings

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

Recent work has managed to learn cross-lingual word embeddings without parallel data by mapping monolingual embeddings to a shared space through adversarial training. However, their evaluation has focused on favorable conditions, using comparable corpora or closely-related languages, and we show that they often fail in more realistic scenarios. This work proposes an alternative approach based on a fully unsupervised initialization that explicitly exploits the structural similarity of the embeddings, and a robust self-learning algorithm that iteratively improves this solution. Our method succeeds in all tested scenarios and obtains the best published results in standard datasets, even surpassing previous supervised systems. Our implementation is released as an open source project at https://github.com/artetxem/vecmap.

1 Introduction

Cross-lingual embedding mappings have shown to be an effective way to learn bilingual word embeddings (Mikolov et al., 2013; Lazaridou et al., 2015). The underlying idea is to independently train the embeddings in different languages using monolingual corpora, and then map them to a shared space through a linear transformation. This allows to learn high-quality cross-lingual representations without expensive supervision, opening new research avenues like unsupervised neural machine translation (Artetxe et al., 2018b; Lample et al., 2018).

While most embedding mapping methods rely on a small seed dictionary, adversarial training has recently produced exciting results in fully unsupervised settings (Zhang et al., 2017a,b; Conneau et al., 2018). However, their evaluation has focused on particularly favorable conditions, limited to closely-related languages or comparable Wikipedia corpora. When tested on more realistic scenarios, we find that they often fail to produce meaningful results. For instance, none of the existing methods works in the standard English-Finnish dataset from Artetxe et al. (2017), obtaining translation accuracies below 2% in all cases (see Section 5).

On another strand of work, Artetxe et al. (2017) showed that an iterative self-learning method is able to bootstrap a high quality mapping from very small seed dictionaries (as little as 25 pairs of words). However, their analysis reveals that the self-learning method gets stuck in poor local optima when the initial solution is not good enough, thus failing for smaller training dictionaries.

In this paper, we follow this second approach and propose a new unsupervised method to build an initial solution without the need of a seed dictionary, based on the observation that, given the similarity matrix of all words in the vocabulary, each word has a different distribution of similarity values. Two equivalent words in different languages should have a similar distribution, and we can use this fact to induce the initial set of word pairings (see Figure 1). We combine this initialization with a more robust self-learning method, which is able to start from the weak initial solution and iteratively improve the mapping. Coupled together, we provide a fully unsupervised cross-lingual mapping method that is effective in realistic settings, converges to a good solution in all cases tested, and sets a new state-of-the-art in bilingual lexicon extraction, even surpassing previous supervised methods.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 789–798
Melbourne, Australia, July 15 - 20, 2018.

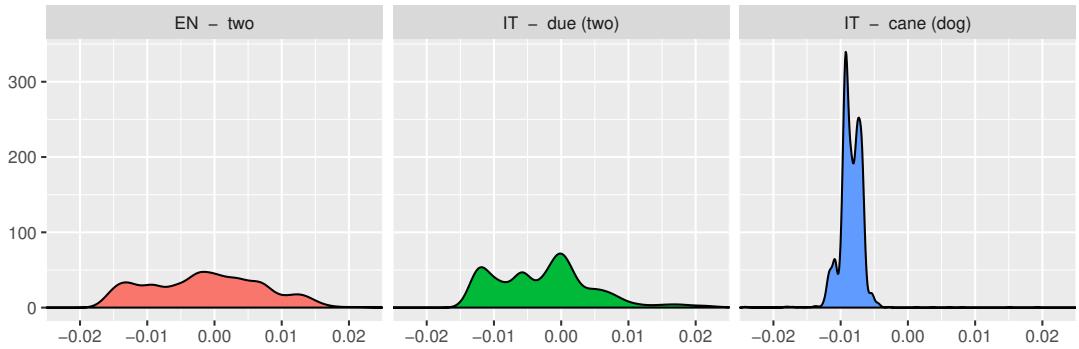

Figure 1: Motivating example for our unsupervised initialization method, showing the similarity distributions of three words (corresponding to the smoothed density estimates from the normalized square root of the similarity matrices as defined in Section 3.2). Equivalent translations (two and *due*) have more similar distributions than non-related words (two and *cane* - meaning dog). This observation is used to build an initial solution that is later improved through self-learning.

2 Related work

Cross-lingual embedding mapping methods work by independently training word embeddings in two languages, and then mapping them to a shared space using a linear transformation.

Most of these methods are **supervised**, and use a bilingual dictionary of a few thousand entries to learn the mapping. Existing approaches can be classified into regression methods, which map the embeddings in one language using a least-squares objective (Mikolov et al., 2013; Shigeto et al., 2015; Dinu et al., 2015), canonical methods, which map the embeddings in both languages to a shared space using canonical correlation analysis and extensions of it (Faruqui and Dyer, 2014; Lu et al., 2015), orthogonal methods, which map the embeddings in one or both languages under the constraint of the transformation being orthogonal (Xing et al., 2015; Artetxe et al., 2016; Zhang et al., 2016; Smith et al., 2017), and margin methods, which map the embeddings in one language to maximize the margin between the correct translations and the rest of the candidates (Lazaridou et al., 2015). Artetxe et al. (2018a) showed that many of them could be generalized as part of a multi-step framework of linear transformations.

A related research line is to adapt these methods to the **semi-supervised** scenario, where the training dictionary is much smaller and used as part of a bootstrapping process. While similar ideas where already explored for traditional count-based vector space models (Peirsman and Padó, 2010; Vulić and Moens, 2013), Artetxe et al. (2017) brought this approach to pre-trained low-dimensional word embeddings, which are more widely used nowadays. More concretely, they proposed a self-learning approach that alternates the mapping and dictionary induction steps iteratively, obtaining results that are comparable to those of supervised methods when starting with only 25 word pairs.

A practical approach for reducing the need of bilingual supervision is to design **heuristics to build the seed dictionary**. The role of the seed lexicon in learning cross-lingual embedding mappings is analyzed in depth by Vulić and Korhonen (2016), who propose using document-aligned corpora to extract the training dictionary. A more common approach is to rely on shared words and cognates (Peirsman and Padó, 2010; Smith et al., 2017), while Artetxe et al. (2017) go further and restrict themselves to shared numerals. However, while these approaches are meant to eliminate the need of bilingual data in practice, they also make strong assumptions on the writing systems of languages (e.g. that they all use a common alphabet or Arabic numerals). Closer to our work, a recent line of **fully unsupervised** approaches drops these assumptions completely, and attempts to learn cross-lingual embedding mappings based on distributional information alone. For that purpose, existing methods rely on adversarial training. This was first proposed by Miceli Barone (2016), who combine an encoder that maps source language embeddings into the target language, a decoder that reconstructs the source language embeddings from the mapped embeddings, and a discriminator that discriminates between the mapped embeddings and the true target language embed-

dings. Despite promising, they conclude that their model "is not competitive with other cross-lingual representation approaches". Zhang et al. (2017a) use a very similar architecture, but incorporate additional techniques like noise injection to aid training and report competitive results on bilingual lexicon extraction. Conneau et al. (2018) drop the reconstruction component, regularize the mapping to be orthogonal, and incorporate an iterative refinement process akin to self-learning, reporting very strong results on a large bilingual lexicon extraction dataset. Finally, Zhang et al. (2017b) adopt the earth mover's distance for training, optimized through a Wasserstein generative adversarial network followed by an alternating optimization procedure. However, all this previous work used comparable Wikipedia corpora in most experiments and, as shown in Section 5, face difficulties in more challenging settings.

3 Proposed method

Let X and Z be the word embedding matrices in two languages, so that their ith row X_{i*} and Z_{i*} denote the embeddings of the ith word in their respective vocabularies. Our goal is to learn the linear transformation matrices W_X and W_Z so the mapped embeddings XW_X and ZW_Z are in the same cross-lingual space. At the same time, we aim to build a dictionary between both languages, encoded as a sparse matrix D where $D_{ij} = 1$ if the jth word in the target language is a translation of the ith word in the source language.

Our proposed method consists of four sequential steps: a pre-processing that normalizes the embeddings (§3.1), a fully unsupervised initialization scheme that creates an initial solution (§3.2), a robust self-learning procedure that iteratively improves this solution (§3.3), and a final refinement step that further improves the resulting mapping through symmetric re-weighting (§3.4).

3.1 Embedding normalization

Our method starts with a pre-processing that length normalizes the embeddings, then mean centers each dimension, and then length normalizes them again. The first two steps have been shown to be beneficial in previous work (Artetxe et al., 2016), while the second length normalization guarantees the final embeddings to have a unit length. As a result, the dot product of any two embeddings is equivalent to their cosine similarity and directly related to their Euclidean distance[1], and can be taken as a measure of their similarity.

3.2 Fully unsupervised initialization

The underlying difficulty of the mapping problem in its unsupervised variant is that the word embedding matrices X and Z are unaligned across both axes: neither the ith vocabulary item X_{i*} and Z_{i*} nor the jth dimension of the embeddings X_{*j} and Z_{*j} are aligned, so there is no direct correspondence between both languages. In order to overcome this challenge and build an initial solution, we propose to first construct two alternative representations X' and Z' that are aligned across their jth dimension X'_{*j} and Z'_{*j}, which can later be used to build an initial dictionary that aligns their respective vocabularies.

Our approach is based on a simple idea: while the axes of the original embeddings X and Z are different in nature, both axes of their corresponding similarity matrices $M_X = XX^T$ and $M_Z = ZZ^T$ correspond to words, which can be exploited to reduce the mismatch to a single axis. More concretely, assuming that the embedding spaces are perfectly isometric, the similarity matrices M_X and M_Z would be equivalent up to a permutation of their rows and columns, where the permutation in question defines the dictionary across both languages. In practice, the isometry requirement will not hold exactly, but it can be assumed to hold approximately, as the very same problem of mapping two embedding spaces without supervision would otherwise be hopeless. Based on that, one could try every possible permutation of row and column indices to find the best match between M_X and M_Z, but the resulting combinatorial explosion makes this approach intractable.

In order to overcome this problem, we propose to first sort the values in each row of M_X and M_Z, resulting in matrices $\mathrm{sorted}(M_X)$ and $\mathrm{sorted}(M_Z)$[2]. Under the strict isometry condition, equivalent words would get the exact same vector across languages, and thus, given a word and its row in $\mathrm{sorted}(M_X)$, one could apply nearest neighbor retrieval over the rows of $\mathrm{sorted}(M_Z)$ to find its corresponding translation.

On a final note, given the singular value decomposition $X = USV^T$, the similarity matrix

[1] Given two length normalized vectors u and v, $u \cdot v = \cos(u, v) = 1 - ||u - v||^2/2$.

[2] Note that the values in each row are sorted independently from other rows.

is $M_X = US^2U^T$. As such, its square root $\sqrt{M_X} = USU^T$ is closer in nature to the original embeddings, and we also find it to work better in practice. We thus compute $\text{sorted}(\sqrt{M_X})$ and $\text{sorted}(\sqrt{M_Z})$ and normalize them as described in Section 3.1, yielding the two matrices X' and Z' that are later used to build the initial solution for self-learning (see Section 3.3).

In practice, the isometry assumption is strong enough so the above procedure captures some cross-lingual signal. In our English-Italian experiments, the average cosine similarity across the gold standard translation pairs is 0.009 for a random solution, 0.582 for the optimal supervised solution, and 0.112 for the mapping resulting from this initialization. While the latter is far from being useful on its own (the accuracy of the resulting dictionary is only 0.52%), it is substantially better than chance, and it works well as an initial solution for the self-learning method described next.

3.3 Robust self-learning

Previous work has shown that self-learning can learn high-quality bilingual embedding mappings starting with as little as 25 word pairs (Artetxe et al., 2017). In this method, training iterates through the following two steps until convergence:

1. Compute the optimal orthogonal mapping maximizing the similarities for the current dictionary D:

$$\arg\max_{W_X, W_Z} \sum_i \sum_j D_{ij}((X_{i*}W_X) \cdot (Z_{j*}W_Z))$$

   An optimal solution is given by $W_X = U$ and $W_Z = V$, where $USV^T = X^TDZ$ is the singular value decomposition of X^TDZ.

2. Compute the optimal dictionary over the similarity matrix of the mapped embeddings $XW_XW_Z^TZ^T$. This typically uses nearest neighbor retrieval from the source language into the target language, so $D_{ij} = 1$ if $j = \text{argmax}_k\,(X_{i*}W_X) \cdot (Z_{k*}W_Z)$ and $D_{ij} = 0$ otherwise.

The underlying optimization objective is independent from the initial dictionary, and the algorithm is guaranteed to converge to a local optimum of it. However, the method does not work if starting from a completely random solution, as it tends to get stuck in poor local optima in that case. For that reason, we use the unsupervised initialization procedure at Section 3.2 to build an initial solution. However, simply plugging in both methods did not work in our preliminary experiments, as the quality of this initial method is not good enough to avoid poor local optima. For that reason, we next propose some key improvements in the dictionary induction step to make self-learning more robust and learn better mappings:

- **Stochastic dictionary induction**. In order to encourage a wider exploration of the search space, we make the dictionary induction stochastic by randomly keeping some elements in the similarity matrix with probability p and setting the remaining ones to 0. As a consequence, the smaller the value of p is, the more the induced dictionary will vary from iteration to iteration, thus enabling to escape poor local optima. So as to find a fine-grained solution once the algorithm gets into a good region, we increase this value during training akin to simulated annealing, starting with $p = 0.1$ and doubling this value every time the objective function at step 1 above does not improve more than $\epsilon = 10^{-6}$ for 50 iterations.

- **Frequency-based vocabulary cutoff**. The size of the similarity matrix grows quadratically with respect to that of the vocabularies. This does not only increase the cost of computing it, but it also makes the number of possible solutions grow exponentially[3], presumably making the optimization problem harder. Given that less frequent words can be expected to be noisier, we propose to restrict the dictionary induction process to the k most frequent words in each language, where we find $k = 20,000$ to work well in practice.

- **CSLS retrieval**. Dinu et al. (2015) showed that nearest neighbor suffers from the hubness problem. This phenomenon is known to occur as an effect of the curse of dimensionality, and causes a few points (known as *hubs*) to be nearest neighbors of many other points (Radovanović et al., 2010a,b). Among the existing solutions to penalize the similarity score of hubs, we adopt the Cross-domain

[3] There are m^n possible combinations that go from a source vocabulary of n entries to a target vocabulary of m entries.

Similarity Local Scaling (CSLS) from Conneau et al. (2018). Given two mapped embeddings x and y, the idea of CSLS is to compute $\mathrm{r_T}(x)$ and $\mathrm{r_S}(y)$, the average cosine similarity of x and y for their k nearest neighbors in the other language, respectively. Having done that, the corrected score $\mathrm{CSLS}(x, y) = 2\cos(x, y) - \mathrm{r_T}(x) - \mathrm{r_S}(y)$. Following the authors, we set $k = 10$.

- **Bidirectional dictionary induction**. When the dictionary is induced from the source into the target language, not all target language words will be present in it, and some will occur multiple times. We argue that this might accentuate the problem of local optima, as repeated words might act as strong attractors from which it is difficult to escape. In order to mitigate this issue and encourage diversity, we propose inducing the dictionary in both directions and taking their corresponding concatenation, so $D = D_{X \to Z} + D_{Z \to X}$.

In order to build the **initial dictionary**, we compute X' and Z' as detailed in Section 3.2 and apply the above procedure over them. As the only difference, this first solution does not use the stochastic zeroing in the similarity matrix, as there is no need to encourage diversity (X' and Z' are only used once), and the threshold for vocabulary cutoff is set to $k = 4,000$, so X' and Z' can fit in memory. Having computed the initial dictionary, X' and Z' are discarded, and the remaining iterations are performed over the original embeddings X and Z.

3.4 Symmetric re-weighting

As part of their multi-step framework, Artetxe et al. (2018a) showed that re-weighting the target language embeddings according to the cross-correlation in each component greatly improved the quality of the induced dictionary. Given the singular value decomposition $USV^T = X^T DZ$, this is equivalent to taking $W_X = U$ and $W_Z = VS$, where X and Z are previously whitened applying the linear transformations $(X^T X)^{-\frac{1}{2}}$ and $(Z^T Z)^{-\frac{1}{2}}$, and later de-whitened applying $U^T (X^T X)^{\frac{1}{2}} U$ and $V^T (Z^T Z)^{\frac{1}{2}} V$.

However, re-weighting also accentuates the problem of local optima when incorporated into self-learning as, by increasing the relevance of dimensions that best match for the current solution, it discourages to explore other regions of the search space. For that reason, we propose using it as a final step once self-learning has converged to a good solution. Unlike Artetxe et al. (2018a), we apply re-weighting symmetrically in both languages, taking $W_X = US^{\frac{1}{2}}$ and $W_Z = VS^{\frac{1}{2}}$. This approach is neutral in the direction of the mapping, and gives good results as shown in our experiments.

4 Experimental settings

Following common practice, we evaluate our method on **bilingual lexicon extraction**, which measures the accuracy of the induced dictionary in comparison to a gold standard.

As discussed before, **previous evaluation** has focused on favorable conditions. In particular, existing unsupervised methods have almost exclusively been tested on Wikipedia corpora, which is comparable rather than monolingual, exposing a strong cross-lingual signal that is not available in strictly unsupervised settings. In addition to that, some datasets comprise unusually small embeddings, with only 50 dimensions and around 5,000-10,000 vocabulary items (Zhang et al., 2017a,b). As the only exception, Conneau et al. (2018) report positive results on the English-Italian dataset of Dinu et al. (2015) in addition to their main experiments, which are carried out in Wikipedia. While this dataset does use strictly monolingual corpora, it still corresponds to a pair of two relatively close indo-european languages.

In order to get a wider picture of how our method compares to previous work in different conditions, including more challenging settings, we carry out our experiments in the widely used **dataset** of Dinu et al. (2015) and the subsequent extensions of Artetxe et al. (2017, 2018a), which together comprise English-Italian, English-German, English-Finnish and English-Spanish. More concretely, the dataset consists of 300-dimensional CBOW embeddings trained on WacKy crawling corpora (English, Italian, German), Common Crawl (Finnish) and WMT News Crawl (Spanish). The gold standards were derived from dictionaries built from Europarl word alignments and available at OPUS (Tiedemann, 2012), split in a test set of 1,500 entries and a training set of 5,000 that we do not use in our experiments. The datasets are freely available. As a non-european agglutinative language, the English-Finnish pair is particularly challeng-

	ES-EN				IT-EN				TR-EN			
	best	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**
Zhang et al. (2017a), $\lambda = 1$	71.43	68.18	**10**	13.2	60.38	56.45	**10**	12.3	0.00	0.00	0	13.0
Zhang et al. (2017a), $\lambda = 10$	70.24	66.37	**10**	13.0	57.64	52.60	**10**	12.6	21.07	17.95	**10**	13.2
Conneau et al. (2018), code	76.18	75.82	**10**	25.1	**67.32**	**67.00**	**10**	25.9	32.64	14.34	5	25.3
Conneau et al. (2018), paper	76.15	75.81	**10**	25.1	67.21	60.22	9	25.5	29.79	16.48	7	25.5
Proposed method	**76.43**	**76.28**	**10**	**0.6**	66.96	66.92	**10**	**0.9**	**36.10**	**35.93**	**10**	**1.7**

Table 1: Results of unsupervised methods on the dataset of Zhang et al. (2017a). We perform 10 runs for each method and report the best and average accuracies (%), the number of successful runs (those with >5% accuracy) and the average runtime (minutes).

	EN-IT				EN-DE				EN-FI				EN-ES			
	best	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**
Zhang et al. (2017a), $\lambda = 1$	0.00	0.00	0	47.0	0.00	0.00	0	47.0	0.00	0.00	0	45.4	0.00	0.00	0	44.3
Zhang et al. (2017a), $\lambda = 10$	0.00	0.00	0	46.6	0.00	0.00	0	46.0	0.07	0.01	0	44.9	0.07	0.01	0	43.0
Conneau et al. (2018), code	45.40	13.55	3	46.1	47.27	42.15	9	45.4	1.62	0.38	0	44.4	36.20	21.23	6	45.3
Conneau et al. (2018), paper	45.27	9.10	2	45.4	0.07	0.01	0	45.0	0.07	0.01	0	44.7	35.47	7.09	2	44.9
Proposed method	**48.53**	**48.13**	**10**	**8.9**	**48.47**	**48.19**	**10**	**7.3**	**33.50**	**32.63**	**10**	**12.9**	**37.60**	**37.33**	**10**	**9.1**

Table 2: Results of unsupervised methods on the dataset of Dinu et al. (2015) and the extensions of Artetxe et al. (2017, 2018a). We perform 10 runs for each method and report the best and average accuracies (%), the number of successful runs (those with >5% accuracy) and the average runtime (minutes).

ing due to the linguistic distance between them. For completeness, we also test our method in the Spanish-English, Italian-English and Turkish-English datasets of Zhang et al. (2017a), which consist of 50-dimensional CBOW embeddings trained on Wikipedia, as well as gold standard dictionaries[4] from Open Multilingual WordNet (Spanish-English and Italian-English) and Google Translate (Turkish-English). The lower dimensionality and comparable corpora make an easier scenario, although it also contains a challenging pair of distant languages (Turkish-English).

Our method is implemented in Python using NumPy and CuPy. Together with it, we also test the **methods** of Zhang et al. (2017a) and Conneau et al. (2018) using the publicly available implementations from the authors[5]. Given that Zhang et al. (2017a) report using a different value of their hyperparameter λ for different language pairs ($\lambda = 10$ for English-Turkish and $\lambda = 1$ for the rest), we test both values in all our experiments to better understand its effect. In the case of Conneau et al. (2018), we test both the default hyperparameters in the source code as well as those reported in the paper, with iterative refinement activated in both cases. Given the instability of these methods, we perform 10 runs for each, and report the best and average accuracies, the number of successful runs (those with >5% accuracy) and the average runtime. All the experiments were run in a single Nvidia Titan Xp.

5 Results and discussion

We first present the main results (§5.1), then the comparison to the state-of-the-art (§5.2), and finally ablation tests to measure the contribution of each component (§5.3).

5.1 Main results

We report the results in the dataset of Zhang et al. (2017a) at Table 1. As it can be seen, the proposed method performs at par with that of Conneau et al. (2018) both in Spanish-English and Italian-English, but gets substantially better results in the more challenging Turkish-English pair. While we are able to reproduce the results reported by Zhang et al. (2017a), their method gets the worst results of all by a large margin. Another disadvantage of that model is that different

[4]The test dictionaries were obtained through personal communication with the authors. The rest of the language pairs were left out due to licensing issues.

[5]Despite our efforts, Zhang et al. (2017b) was left out because: 1) it does not create a one-to-one dictionary, thus difficulting direct comparison, 2) it depends on expensive proprietary software 3) its computational cost is orders of magnitude higher (running the experiments would have taken several months).

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (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 et al. (2015)	37.7	38.93*	29.14*	30.40*
	Lazaridou et al. (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 et al. (2016)	39.27	41.87*	30.62*	31.40*
	Artetxe et al. (2017)	39.67	40.87	28.72	-
	Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
	Artetxe et al. (2018a)	45.27	44.13	**32.94**	36.60
25 dict.	Artetxe et al. (2017)	37.27	39.60	28.16	-
Init. heurist.	Smith et al. (2017), cognates	39.9	-	-	-
	Artetxe et al. (2017), num.	39.40	40.27	26.47	-
None	Zhang et al. (2017a), $\lambda = 1$	0.00*	0.00*	0.00*	0.00*
	Zhang et al. (2017a), $\lambda = 10$	0.00*	0.00*	0.01*	0.01*
	Conneau et al. (2018), code‡	45.15*	46.83*	0.38*	35.38*
	Conneau et al. (2018), paper‡	45.1	0.01*	0.01*	35.44*
	Proposed method	**48.13**	**48.19**	32.63	**37.33**

Table 3: Accuracy (%) of the proposed method in comparison with previous work. *Results obtained with the official implementation from the authors. †Results obtained with the framework from Artetxe et al. (2018a). The remaining results were reported in the original papers. For methods that do not require supervision, we report the average accuracy across 10 runs. ‡For meaningful comparison, runs with <5% accuracy are excluded when computing the average, but note that, unlike ours, their method often gives a degenerated solution (see Table 2).

language pairs require different hyperparameters: $\lambda = 1$ works substantially better for Spanish-English and Italian-English, but only $\lambda = 10$ works for Turkish-English.

The results for the more challenging dataset from Dinu et al. (2015) and the extensions of Artetxe et al. (2017, 2018a) are given in Table 2. In this case, our proposed method obtains the best results in all metrics for all the four language pairs tested. The method of Zhang et al. (2017a) does not work at all in this more challenging scenario, which is in line with the negative results reported by the authors themselves for similar conditions (only %2.53 accuracy in their large Gigaword dataset). The method of Conneau et al. (2018) also fails for English-Finnish (only 1.62% in the best run), although it is able to get positive results in some runs for the rest of language pairs. Between the two configurations tested, the default hyperparameters in the code show a more stable behavior.

These results confirm the robustness of the proposed method. While the other systems succeed in some runs and fail in others, our method converges to a good solution in all runs without exception and, in fact, it is the only one getting positive results for English-Finnish. In addition to being more robust, our method also obtains substantially better accuracies, surpassing previous methods by at least 1-3 points in all but the easiest pairs. Moreover, our method is not sensitive to hyperparameters that are difficult to tune without a development set, which is critical in realistic unsupervised conditions.

At the same time, our method is significantly faster than the rest. In relation to that, it is interesting that, while previous methods perform a fixed number of iterations and take practically the same time for all the different language pairs, the runtime of our method adapts to the difficulty of the task thanks to the dynamic convergence criterion of our stochastic approach. This way, our method tends to take longer for more challenging language pairs (1.7 vs 0.6 minutes for es-en and tr-en in one dataset, and 12.9 vs 7.3 minutes for en-fi and en-de in the other) and, in fact, our (relative) execution times correlate surprisingly well with the linguistic distance with English (closest/fastest is German, followed by Italian/Spanish, followed by Turkish/Finnish).

	EN-IT				EN-DE				EN-FI				EN-ES			
	best	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**	**best**	**avg**	**s**	**t**
Full system	48.53	48.13	10	8.9	48.47	48.19	10	7.3	33.50	32.63	10	12.9	37.60	37.33	10	9.1
- Unsup. init.	0.07	0.02	0	16.5	0.00	0.00	0	17.3	0.07	0.01	0	13.8	0.13	0.02	0	15.9
- Stochastic	48.20	48.20	10	2.7	48.13	48.13	10	2.5	0.28	0.28	0	4.3	37.80	37.80	10	2.6
- Cutoff (k=100k)	46.87	46.46	10	114.5	48.27	48.12	10	105.3	31.95	30.78	10	162.5	35.47	34.88	10	185.2
- CSLS	0.00	0.00	0	15.0	0.00	0.00	0	13.8	0.00	0.00	0	13.1	0.00	0.00	0	14.1
- Bidirectional	46.00	45.37	10	5.6	48.27	48.03	10	5.5	31.39	24.86	8	7.8	36.20	35.77	10	7.3
- Re-weighting	46.07	45.61	10	8.4	48.13	47.41	10	7.0	32.94	31.77	10	11.2	36.00	35.45	10	9.1

Table 4: Ablation test on the dataset of Dinu et al. (2015) and the extensions of Artetxe et al. (2017, 2018a). We perform 10 runs for each method and report the best and average accuracies (%), the number of successful runs (those with >5% accuracy) and the average runtime (minutes).

5.2 Comparison with the state-of-the-art

Table 3 shows the results of the proposed method in comparison to previous systems, including those with different degrees of supervision. We focus on the widely used English-Italian dataset of Dinu et al. (2015) and its extensions. Despite being fully unsupervised, our method achieves the best results in all language pairs but one, even surpassing previous supervised approaches. The only exception is English-Finnish, where Artetxe et al. (2018a) gets marginally better results with a difference of 0.3 points, yet ours is the only unsupervised system that works for this pair. At the same time, it is remarkable that the proposed system gets substantially better results than Artetxe et al. (2017), the only other system based on self-learning, with the additional advantage of being fully unsupervised.

5.3 Ablation test

In order to better understand the role of different aspects in the proposed system, we perform an ablation test, where we separately analyze the effect of initialization, the different components of our robust self-learning algorithm, and the final symmetric re-weighting. The obtained results are reported in Table 4.

In concordance with previous work, our results show that self-learning does not work with random initialization. However, the proposed unsupervised initialization is able to overcome this issue without the need of any additional information, performing at par with other character-level heuristics that we tested (e.g. shared numerals).

As for the different self-learning components, we observe that the stochastic dictionary induction is necessary to overcome the problem of poor local optima for English-Finnish, although it does not make any difference for the rest of easier language pairs. The frequency-based vocabulary cutoff also has a positive effect, yielding to slightly better accuracies and much faster runtimes. At the same time, CSLS plays a critical role in the system, as hubness severely accentuates the problem of local optima in its absence. The bidirectional dictionary induction is also beneficial, contributing to the robustness of the system as shown by English-Finnish and yielding to better accuracies in all cases.

Finally, these results also show that symmetric re-weighting contributes positively, bringing an improvement of around 1-2 points without any cost in the execution time.

6 Conclusions

In this paper, we show that previous unsupervised mapping methods (Zhang et al., 2017a; Conneau et al., 2018) often fail on realistic scenarios involving non-comparable corpora and/or distant languages. In contrast to adversarial methods, we propose to use an initial weak mapping that exploits the structure of the embedding spaces in combination with a robust self-learning approach. The results show that our method succeeds in all cases, providing the best results with respect to all previous work on unsupervised and supervised mappings.

The ablation analysis shows that our initial solution is instrumental for making self-learning work without supervision. In order to make self-learning robust, we also added stochasticity to dictionary induction, used CSLS instead of nearest neighbor, and produced bidirectional dictionaries. Results also improved using smaller in-

termediate vocabularies and re-weighting the final solution. Our implementation is available as an open source project at https://github.com/artetxem/vecmap.

In the future, we would like to extend the method from the bilingual to the multilingual scenario, and go beyond the word level by incorporating embeddings of longer phrases.

Acknowledgments

This research was partially supported by the Spanish MINECO (TUNER TIN2015-65308-C5-1-R, MUSTER PCIN-2015-226 and TADEEP TIN2015-70214-P, cofunded by EU FEDER), the UPV/EHU (excellence research group), and the NVIDIA GPU grant program. Mikel Artetxe enjoys a doctoral grant from the Spanish MECD.

References

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, Austin, Texas. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. 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)*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

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

Manaal Faruqui and Chris Dyer. 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*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 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)*, pages 270–280, Beijing, China. Association for Computational Linguistics.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 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*, pages 250–256, Denver, Colorado. Association for Computational Linguistics.

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

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

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929, Los Angeles, California. Association for Computational Linguistics.

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

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 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*, pages 186–193.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. 2015. Ridge regression, hubness, and zero-shot learning. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Proceedings, Part I*, pages 135–151.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ivan Vulić and Anna Korhonen. 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)*, pages 247–257, Berlin, Germany. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1613–1624, Seattle, Washington, USA. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 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*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 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*, pages 1307–1317, San Diego, California. Association for Computational Linguistics.