Learning bilingual word embeddings with (almost) no bilingual data

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

Abstract

Most methods to learn bilingual word embeddings rely on large parallel corpora, which is difficult to obtain for most language pairs. This has motivated an active research line to relax this requirement, with methods that use document-aligned corpora or bilingual dictionaries of a few thousand words instead. In this work, we further reduce the need of bilingual resources using a very simple self-learning approach that can be combined with any dictionary-based mapping technique. Our method exploits the structural similarity of embedding spaces, and works with as little bilingual evidence as a 25 word dictionary or even an automatically generated list of numerals, obtaining results comparable to those of systems that use richer resources.

1 Introduction

Multilingual word embeddings have attracted a lot of attention in recent times. In addition to having a direct application in inherently crosslingual tasks like machine translation (Zou et al., 2013) and crosslingual entity linking (Tsai and Roth, 2016), they provide an excellent mechanism for transfer learning, where a model trained in a resource-rich language is transferred to a less-resourced one, as shown with part-of-speech tagging (Zhang et al., 2016), parsing (Xiao and Guo, 2014) and document classification (Klementiev et al., 2012).

Most methods to learn these multilingual word embeddings make use of large parallel corpora (Gouws et al., 2015; Luong et al., 2015), but there have been several proposals to relax this requirement, given its scarcity in most language pairs. A possible relaxation is to use document-aligned or label-aligned comparable corpora (Søgaard et al., 2015; Vulić and Moens, 2016; Mogadala and Rettinger, 2016), but large amounts of such corpora are not always available for some language pairs.

An alternative approach that we follow here is to independently train the embeddings for each language on monolingual corpora, and then learn a linear transformation to map the embeddings from one space into the other by minimizing the distances in a bilingual dictionary, usually in the range of a few thousand entries (Mikolov et al., 2013a; Artetxe et al., 2016). However, dictionaries of that size are not readily available for many language pairs, specially those involving less-resourced languages.

In this work, we reduce the need of large bilingual dictionaries to much smaller seed dictionaries. Our method can work with as little as 25 word pairs, which are straightforward to obtain assuming some basic knowledge of the languages involved. The method can also work with trivially generated seed dictionaries of numerals (i.e. 1-1, 2-2, 3-3, 4-4...) making it possible to learn bilingual word embeddings without any real bilingual data. In either case, we obtain very competitive results, comparable to other state-of-the-art methods that make use of much richer bilingual resources.

The proposed method is an extension of existing mapping techniques, where the dictionary is used to learn the embedding mapping and the embedding mapping is used to induce a new dictionary iteratively in a self-learning fashion (see Figure 1). In spite of its simplicity, our analysis of the implicit optimization objective reveals that the method is exploiting the structural similarity of independently trained embeddings.

We analyze previous work in Section 2. Section 3 describes the self-learning framework, while Section 4 presents the experiments. Section 5 analyzes the underlying optimization objective, and Section 6 presents an error analysis.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 451–462
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1042

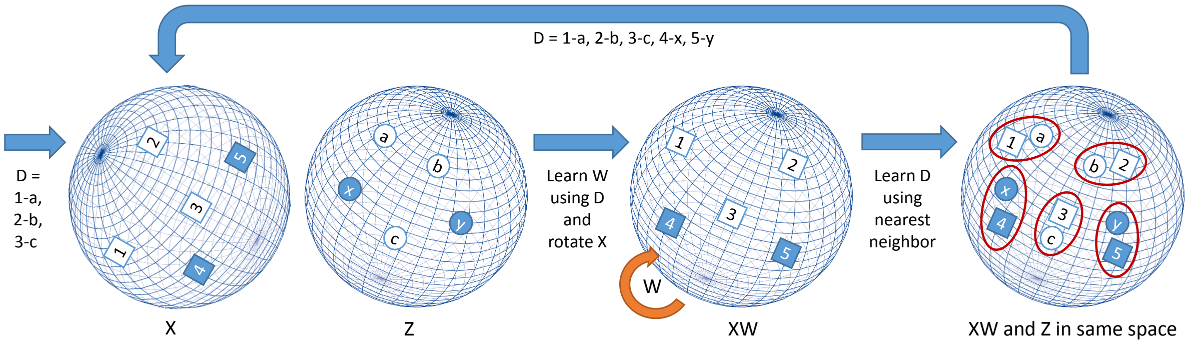

Figure 1: A general schema of the proposed self-learning framework. Previous works learn a mapping W based on the seed dictionary D, which is then used to learn the full dictionary. In our proposal we use the new dictionary to learn a new mapping, iterating until convergence.

2 Related work

We will first focus on bilingual embedding mappings, which are the basis of our proposals, and then on other unsupervised and weakly supervised methods to learn bilingual word embeddings.

2.1 Bilingual embedding mappings

Methods to induce bilingual mappings work by independently learning the embeddings in each language using monolingual corpora, and then learning a transformation from one embedding space into the other based on a bilingual dictionary.

The first of such methods is due to Mikolov et al. (2013a), who learn the linear transformation that minimizes the sum of squared Euclidean distances for the dictionary entries. The same optimization objective is used by Zhang et al. (2016), who constrain the transformation matrix to be orthogonal. Xing et al. (2015) incorporate length normalization in the training of word embeddings and maximize the cosine similarity instead, enforcing the orthogonality constraint to preserve the length normalization after the mapping. Finally, Lazaridou et al. (2015) use max-margin optimization with intruder negative sampling.

Instead of learning a single linear transformation from the source language into the target language, Faruqui and Dyer (2014) use canonical correlation analysis to map both languages to a shared vector space. Lu et al. (2015) extend this work and apply deep canonical correlation analysis to learn non-linear transformations.

Artetxe et al. (2016) propose a general framework that clarifies the relation between Mikolov et al. (2013a), Xing et al. (2015), Faruqui and Dyer (2014) and Zhang et al. (2016) as variants of the same core optimization objective, and show that a new variant is able to surpass them all. While most of the previous methods use gradient descent, Artetxe et al. (2016) propose an efficient analytical implementation for those same methods, recently extended by Smith et al. (2017) to incorporate dimensionality reduction.

A prominent application of bilingual embedding mappings, with a direct application in machine translation (Zhao et al., 2015), is bilingual lexicon extraction, which is also the main evaluation method. More specifically, the learned mapping is used to induce the translation of source language words that were missing in the original dictionary, usually by taking their nearest neighbor word in the target language according to cosine similarity, although Dinu et al. (2015) and Smith et al. (2017) propose alternative retrieval methods to address the hubness problem.

2.2 Unsupervised and weakly supervised bilingual embeddings

As mentioned before, our method works with as little as 25 word pairs, while the methods discussed previously use thousands of pairs. The only exception in this regard is the work by Zhang et al. (2016), who only use 10 word pairs with good results on transfer learning for part-of-speech tagging. Our experiments will show that, although their method captures coarse-grained relations, it fails on finer-grained tasks like bilingual lexicon induction.

Bootstrapping methods similar to ours have been previously proposed for traditional count-based vector space models (Peirsman and Padó, 2010; Vulić and Moens, 2013). However, while previous techniques incrementally build a high-

Algorithm 1 Traditional framework

```
Input: $X$ (source embeddings)
Input: $Z$ (target embeddings)
Input: $D$ (seed dictionary)
$W \leftarrow$ LEARN_MAPPING($X$, $Z$, $D$)
$D \leftarrow$ LEARN_DICTIONARY($X$, $Z$, $W$)
EVALUATE_DICTIONARY($D$)
```

dimensional model where each axis encodes the co-occurrences with a specific word and its equivalent in the other language, our method works with low-dimensional pre-trained word embeddings, which are more widely used nowadays.

A practical aspect for reducing the need of bilingual supervision is on the design of the seed dictionary. This 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), eliminating the need of bilingual data in practice. Our use of shared numerals exploits the same underlying idea, but relies on even less bilingual evidence and should thus generalize better to distant language pairs.

Miceli Barone (2016) and Cao et al. (2016) go one step further and attempt to learn bilingual embeddings without any bilingual evidence. The former uses adversarial autoencoders (Makhzani et al., 2016), combining an encoder that maps the source language embeddings into the target language, a decoder that reconstructs the original embeddings, and a discriminator that distinguishes mapped embeddings from real target language embeddings, whereas the latter adds a regularization term to the training of word embeddings that pushes the mean and variance of each dimension in different languages close to each other. Although promising, the reported performance in both cases is poor in comparison to other methods.

Finally, the induction of bilingual knowledge from monolingual corpora is closely related to the decipherment scenario, for which models that incorporate word embeddings have also been proposed (Dou et al., 2015). However, decipherment is only concerned with translating text from one language to another and relies on complex statistical models that are designed specifically for that purpose, while our approach is more general and learns task-independent multilingual embeddings.

Algorithm 2 Proposed self-learning framework

```
Input: $X$ (source embeddings)
Input: $Z$ (target embeddings)
Input: $D$ (seed dictionary)
repeat
    $W \leftarrow$ LEARN_MAPPING($X$, $Z$, $D$)
    $D \leftarrow$ LEARN_DICTIONARY($X$, $Z$, $W$)
until convergence criterion
EVALUATE_DICTIONARY($D$)
```

3 Proposed self-learning framework

As discussed in Section 2.1, a common evaluation task (and practical application) of bilingual embedding mappings is to induce bilingual lexicons, that is, to obtain the translation of source words that were missing in the training dictionary, which are then compared to a gold standard test dictionary for evaluation. This way, one can say that the seed (train) dictionary is used to learn a mapping, which is then used to induce a better dictionary (at least in the sense that it is larger). Algorithm 1 summarizes this framework.

Following this observation, we propose to use the output dictionary in Algorithm 1 as the input of the same system in a self-learning fashion which, assuming that the output dictionary was indeed better than the original one, should serve to learn a better mapping and, consequently, an even better dictionary the second time. The process can then be repeated iteratively to obtain a hopefully better mapping and dictionary each time until some convergence criterion is met. Algorithm 2 summarizes this alternative framework that we propose.

Our method can be combined with any embedding mapping and dictionary induction technique (see Section 2.1). However, efficiency turns out to be critical for a variety of reasons. First of all, by enclosing the learning logic in a loop, the total training time is increased by the number of iterations. Even more importantly, our framework requires to explicitly build the entire dictionary at each iteration, whereas previous work tends to induce the translation of individual words on-demand later at runtime. Moreover, from the second iteration onwards, it is this induced, full dictionary that has to be used to learn the embedding mapping, and not the considerably smaller seed dictionary as it is typically done. In the following two subsections, we respectively describe the embedding mapping method and the dictionary induction method that we adopt in our work with these efficiency requirements in mind.

3.1 Embedding mapping

As discussed in Section 2.1, most previous methods to learn embedding mappings use variants of gradient descent. Among the more efficient exact alternatives, we decide to adopt the one by Artetxe et al. (2016) for its simplicity and good results as reported in their paper. We next present their method, adapting the formalization to explicitly incorporate the dictionary as required by our self-learning algorithm.

Let X and Z denote the word embedding matrices in two languages so that X_{i*} corresponds to the ith source language word embedding and Z_{j*} corresponds to the jth target language embedding. While Artetxe et al. (2016) assume these two matrices are aligned according to the dictionary, we drop this assumption and represent the dictionary explicitly as a binary matrix D, so that $D_{ij} = 1$ if the ith source language word is aligned with the jth target language word. The goal is then to find the optimal mapping matrix W^* so that the sum of squared Euclidean distances between the mapped source embeddings $X_{i*}W$ and target embeddings Z_{j*} for the dictionary entries D_{ij} is minimized:

$$W^* = \arg\min_W \sum_i \sum_j D_{ij} ||X_{i*}W - Z_{j*}||^2$$

Following Artetxe et al. (2016), we length normalize and mean center the embedding matrices X and Z in a preprocessing step, and constrain W to be an orthogonal matrix (i.e. $WW^T = W^TW = I$), which serves to enforce monolingual invariance, preventing a degradation in monolingual performance while yielding to better bilingual mappings. Under such orthogonality constraint, minimizing the squared Euclidean distance becomes equivalent to maximizing the dot product, so the above optimization objective can be reformulated as follows:

$$W^* = \arg\max_W \operatorname{Tr}\left(XWZ^TD^T\right)$$

where $\operatorname{Tr}(\cdot)$ denotes the trace operator (the sum of all the elements in the main diagonal). The optimal orthogonal solution for this problem is given by $W^* = UV^T$, where $X^TDZ = U\Sigma V^T$ is the singular value decomposition of X^TDZ. Since the dictionary matrix D is sparse, this can be efficiently computed in linear time with respect to the number of dictionary entries.

3.2 Dictionary induction

As discussed in Section 2.1, practically all previous work uses nearest neighbor retrieval for word translation induction based on embedding mappings. In nearest neighbor retrieval, each source language word is assigned the closest word in the target language. In our work, we use the dot product between the mapped source language embeddings and the target language embeddings as the similarity measure, which is roughly equivalent to cosine similarity given that we apply length normalization followed by mean centering as a preprocessing step (see Section 3.1). This way, following the notation in Section 3.1, we set $D_{ij} = 1$ if $j = \operatorname{argmax}_k (X_{i*}W) \cdot Z_{k*}$ and $D_{ij} = 0$ otherwise[1].

While we find that independently computing the similarity measure between all word pairs is prohibitively slow, the computation of the entire similarity matrix XWZ^T can be easily vectorized using popular linear algebra libraries, obtaining big performance gains. However, the resulting similarity matrix is often too large to fit in memory when using large vocabularies. For that reason, instead of computing the entire similarity matrix XWZ^T in a single step, we iteratively compute submatrices of it using vectorized matrix multiplication, find their corresponding maxima each time, and then combine the results.

4 Experiments and results

In this section, we experimentally test the proposed method in bilingual lexicon induction and crosslingual word similarity. Subsection 4.1 describes the experimental settings, while Subsections 4.2 and 4.3 present the results obtained in each of the tasks. The code and resources necessary to reproduce our experiments are available at `https://github.com/artetxem/vecmap`.

4.1 Experimental settings

For easier comparison with related work, we evaluated our mappings on **bilingual lexicon induction** using the public **English-Italian** dataset by Dinu et al. (2015), which includes monolingual word embeddings in both languages together with a bilingual dictionary split in a training set and a

[1] Note that we induce the dictionary entries starting from the source language words. We experimented with other alternatives in development, with minor differences.

test set[2]. The embeddings were trained with the word2vec toolkit with CBOW and negative sampling (Mikolov et al., 2013b)[3], using a 2.8 billion word corpus for English (ukWaC + Wikipedia + BNC) and a 1.6 billion word corpus for Italian (itWaC). The training and test sets were derived from a dictionary built form 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 of the remaining word pairs as the training set.

In addition to English-Italian, we selected two other languages from different language families with publicly available resources. We thus created analogous datasets for **English-German** and **English-Finnish**. In the case of German, the embeddings were trained on the 0.9 billion word corpus SdeWaC, which is part of the WaCky collection (Baroni et al., 2009) that was also used for English and Italian. Given that Finnish is not included in this collection, we used the 2.8 billion word Common Crawl corpus provided at WMT 2016[4] instead, which we tokenized using the Stanford Tokenizer (Manning et al., 2014). In addition to that, we created training and test sets for both pairs from their respective Europarl dictionaries from OPUS following the exact same procedure used for English-Italian, and the word embeddings were also trained using the same configuration as Dinu et al. (2015).

Given that the main focus of our work is on **small seed dictionaries**, we created random subsets of 2,500, 1,000, 500, 250, 100, 75, 50 and 25 entries from the original training dictionaries of 5,000 entries. This was done by shuffling once the training dictionaries and taking their first k entries, so it is guaranteed that each dictionary is a strict subset of the bigger dictionaries.

In addition to that, we explored using automatically generated dictionaries as a shortcut to practical unsupervised learning. For that purpose, we created **numeral dictionaries**, consisting of words matching the `[0-9]+` regular expression in both vocabularies (e.g. 1-1, 2-2, 3-3, 1992-1992 etc.). The resulting dictionary had 2772 entries for English-Italian, 2148 for English-German, and 2345 for English-Finnish. While more sophisticated approaches are possible (e.g. involving the edit distance of all words), we believe that this method is general enough that should work with practically any language pair, as Arabic numerals are often used even in languages with a different writing system (e.g. Chinese and Russian).

While bilingual lexicon induction is a standard evaluation task for seed dictionary based methods like ours, it is unsuitable for bilingual corpus based methods, as statistical word alignment already provides a reliable way to derive dictionaries from bilingual corpora and, in fact, this is how the test dictionary itself is built in our case. For that reason, we carried out some experiments in **crosslingual word similarity** as a way to test our method in a different task and allowing to compare it to systems that use richer bilingual data. There are no many crosslingual word similarity datasets, and we used the RG-65 and WordSim-353 crosslingual datasets for English-German and the WordSim-353 crosslingual dataset for English-Italian as published by Camacho-Collados et al. (2015) [5].

As for the **convergence criterion**, we decide to stop training when the improvement on the average dot product for the induced dictionary falls below a given threshold from one iteration to the next. After length normalization, the dot product ranges from -1 to 1, so we decide to set this threshold at 1e-6, which we find to be a very conservative value yet enough that training takes a reasonable amount of time. The curves in the next section confirm that this was a reasonable choice.

This convergence criterion is usually met in less than 100 iterations, each of them taking 5 minutes on a modest desktop computer (Intel Core i5-4670 CPU with 8GiB of RAM), including the induction of a dictionary of 200,000 words at each iteration.

4.2 Bilingual lexicon induction

For the experiments on bilingual lexicon induction, we compared our method with those proposed by Mikolov et al. (2013a), Xing et al. (2015), Zhang et al. (2016) and Artetxe et al. (2016), all of them implemented as part of the framework proposed by the latter. The results ob-

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

[3] 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

[4] `http://www.statmt.org/wmt16/translation-task.html`

[5] `http://lcl.uniroma1.it/similarity-datasets/`

	English-Italian			English-German			English-Finnish		
	5,000	**25**	**num.**	**5,000**	**25**	**num.**	**5,000**	**25**	**num.**
Mikolov et al. (2013a)	34.93	0.00	0.00	35.00	0.00	0.07	25.91	0.00	0.00
Xing et al. (2015)	36.87	0.00	0.13	41.27	0.07	0.53	28.23	0.07	0.56
Zhang et al. (2016)	36.73	0.07	0.27	40.80	0.13	0.87	28.16	0.14	0.42
Artetxe et al. (2016)	39.27	0.07	0.40	**41.87**	0.13	0.73	**30.62**	0.21	0.77
Our method	**39.67**	**37.27**	**39.40**	40.87	**39.60**	**40.27**	28.72	**28.16**	**26.47**

Table 1: Accuracy (%) on bilingual lexicon induction for different seed dictionaries

tained with the 5,000 entry, 25 entry and the numerals dictionaries for all the 3 language pairs are given in Table 1.

The results for the 5,000 entry dictionaries show that our method is comparable or even better than the other systems. As another reference, the best published results using nearest-neighbor retrieval are due to Lazaridou et al. (2015), who report an accuracy of 40.20% for the full English-Italian dictionary, almost at pair with our system (39.67%).

In any case, the main focus of our work is on smaller dictionaries, and it is under this setting that our method really stands out. The 25 entry and numerals columns in Table 1 show the results for this setting, where all previous methods drop dramatically, falling below 1% accuracy in all cases. The method by Zhang et al. (2016) also obtains poor results with small dictionaries, which reinforces our hypothesis in Section 2.2 that their method can only capture coarse-grain bilingual relations for small dictionaries. In contrast, our proposed method obtains very competitive results for all dictionaries, with a difference of only 1-2 points between the full dictionary and both the 25 entry dictionary and the numerals dictionary in all three languages. Figure 2 shows the curve of the English-Italian accuracy for different seed dictionary sizes, confirming this trend.

Finally, it is worth mentioning that, even if all the three language pairs show the same general behavior, there are clear differences in their absolute accuracy numbers, which can be attributed to the linguistic proximity of the languages involved. In particular, the results for English-Finnish are about 10 points below the rest, which is explained by the fact that Finnish is a non-indoeuropean agglutinative language, making the task considerably more difficult for this language pair. In this regard, we believe that the good results with small dictionaries are a strong indication of the robustness of our method, showing that it is able to learn good bilingual mappings from very little bilingual evidence even for distant language pairs where the structural similarity of the embedding spaces is presumably weaker.

4.3 Crosslingual word similarity

In addition to the baseline systems in Section 4.2, in the crosslingual similarity experiments we also tested the method by Luong et al. (2015), which is the state-of-the-art for bilingual word embeddings based on parallel corpora (Upadhyay et al., 2016)[6]. As this method is an extension of word2vec, we used the same hyperparameters as for the monolingual embeddings when possible (see Section 4.1), and leave the default ones otherwise. We used Europarl as our parallel corpus to train this method as done by the authors, which consists of nearly 2 million parallel sentences.

As shown in the results in Table 2, our method obtains the best results in all cases, surpassing the rest of the dictionary-based methods by 1-3 points depending on the dataset. But, most importantly, it does not suffer from any significant degradation for using smaller dictionaries and, in fact, our method gets better results using the 25 entry dictionary or the numeral list as the only bilingual evidence than any of the baseline systems using much richer resources.

The relatively poor results of Luong et al. (2015) can be attributed to the fact that the dictionary based methods make use of much bigger monolingual corpora, while methods based on parallel corpora are restricted to smaller corpora. However, it is not clear how to introduce monolingual corpora on those methods. We did run some experiments with BilBOWA (Gouws et al., 2015), which supports training in monolingual corpora in addition to bilingual corpora, but obtained very poor results[7]. All in all, our experiments show

[6] We also tested English-German pre-trained embeddings from Klementiev et al. (2012) and Chandar A P et al. (2014). They both had coverage problems that made the results hard to compare, and, when considering the correlations for the word pairs in their vocabulary, their performance was poor.

[7] Upadhyay et al. (2016) report similar problems using

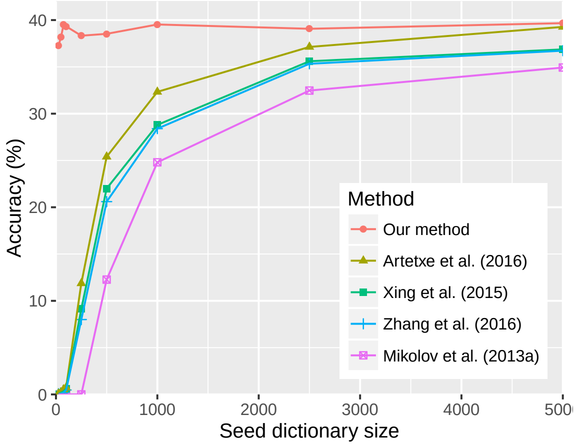

Figure 2: Accuracy on English-Italian bilingual lexicon induction for different seed dictionaries

that it is better to use large monolingual corpora in combination with very little bilingual data rather than a bilingual corpus of a standard size alone.

5 Global optimization objective

It might seem somehow surprising at first that, as seen in the previous section, our simple self-learning approach is able to learn high quality bilingual embeddings from small seed dictionaries instead of falling in degenerated solutions. In this section, we try to shed light on our approach, and give empirical evidence supporting our claim.

More concretely, we argue that, for the embedding mapping and dictionary induction methods described in Section 3, the proposed self-learning framework is implicitly solving the following global optimization problem[8]:

$$W^* = \arg\max_W \sum_i \max_j \left(X_{i*}W\right) \cdot Z_{j*} \qquad \text{s.t.} \quad WW^T = W^TW = I$$

Contrary to the optimization objective for W in Section 3.1, the global optimization objective does not refer to any dictionary, and maximizes the similarity between each source language word and its closest target language word. Intuitively, a random solution would map source language embeddings to seemingly random locations in the target language space, and it would thus be unlikely that they have any target language word nearby, making the optimization value small. In contrast, a good solution would map source language words close to their translation equivalents in the target language space, and they would thus have their corresponding embeddings nearby, making the optimization value large. While it is certainly possible to build degenerated solutions that take high optimization values for small subsets of the vocabulary, we think that the structural similarity between independently trained embedding spaces in different languages is strong enough that optimizing this function yields to meaningful bilingual mappings when the size of the vocabulary is much larger than the dimensionality of the embeddings.

		IT	DE	
	Bi. data	**WS**	**RG**	**WS**
Luong et al. (2015)	Europarl	.331	.335	.356
Mikolov et al. (2013a)	5k dict	.627	.643	.528
Xing et al. (2015)	5k dict	.614	.700	.595
Zhang et al. (2016)	5k dict	.616	.704	.596
Artetxe et al. (2016)	5k dict	.617	.716	.597
Our method	5k dict	.624	.742	**.616**
	25 dict	.626	**.749**	.612
	num.	**.628**	.739	.604

Table 2: Spearman correlations on English-Italian and English-German crosslingual word similarity

The reasoning for how the self-learning framework is optimizing this objective is as follows. At the end of each iteration, the dictionary D is updated to assign, for the current mapping W, each source language word to its closest target language word. This way, when we update W to maximize the average similarity of these dictionary entries at the beginning of the next iteration, it is guaranteed that the value of the optimization objective will improve (or at least remain the same). The reason is that the average similarity between each word and what were previously the closest words will be improved if possible, as this is what the updated W directly optimizes (see Section 3.1). In addition to that, it is also possible that, for some source words, some other target words get closer after the update. Thanks to this, our self-learning algorithm is guaranteed to converge to a local optimum of the above global objective, behaving like an alternating optimization algorithm for it.

It is interesting to note that the above reasoning is valid no matter what the the initial solution is, and, in fact, the global optimization objective does not depend on the seed dictionary nor any other

BilBOWA.

[8] While we restrict our formal analysis to the embedding mapping and dictionary induction method that we use, the general reasoning should be valid for other choices as well.

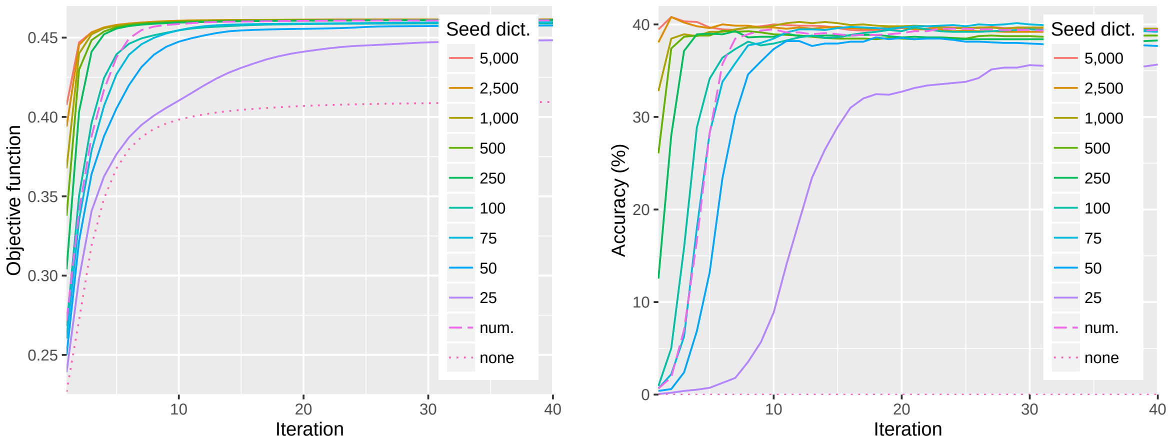

Figure 3: Learning curve on English-Italian according to the global objective function (left) and the accuracy on bilingual lexicon induction (right)

bilingual resource. For that reason, it should be possible to use a random initialization instead of a small seed dictionary. However, we empirically observe that this works poorly in practice, as our algorithm tends to get stuck in poor local optima when the initial solution is not good enough.

The general behavior of our method is reflected in Figure 3, which shows the learning curve for different seed dictionaries according to both the objective function and the accuracy on bilingual lexicon induction. As it can be seen, the objective function is improved from iteration to iteration and converges to a local optimum just as expected. At the same time, the learning curves show a strong correlation between the optimization objective and the accuracy, as it can be clearly observed that improving the former leads to an improvement of the latter, confirming our explanations. Regarding random initialization, the figure shows that the algorithm gets stuck in a poor local optimum of the objective function, which is the reason of the bad performance (0% accuracy) on bilingual lexicon induction, but the proposed optimization objective itself seems to be adequate.

Finally, we empirically observe that our algorithm learns similar mappings no matter what the seed dictionary was. We first repeated our experiments on English-Italian bilingual lexicon induction for 5 different dictionaries of 25 entries, obtaining an average accuracy of 38.15% and a standard deviation of only 0.75%. In addition to that, we observe that the overlap between the predictions made when starting with the full dictionary and the numerals dictionary is 76.00% (60.00% for the 25 entry dictionary). At the same time, 37.00% of the test cases are correctly solved by both instances, and it is only 5.07% of the test cases that one of them gets right and the other wrong (34.00% and 8.94% for the 25 entry dictionary). This suggests that our algorithm tends to converge to similar solutions even for disjoint seed dictionaries, which is in line with our view that we are implicitly optimizing an objective that is independent from the seed dictionary, yet a seed dictionary is necessary to build a good enough initial solution to avoid getting stuck in poor local optima. For that reason, it is likely that better methods to tackle this optimization problem would allow learning bilingual word embeddings without any bilingual evidence at all and, in this regard, we believe that our work opens exciting opportunities for future research.

6 Error analysis

So as to better understand the behavior of our system, we performed an error analysis of its output in English-Italian bilingual lexicon induction when starting with the 5,000 entry, the 25 entry and the numeral dictionaries in comparison with the baseline method of Artetxe et al. (2016) with the 5,000 entry dictionary. For that purpose, we took 100 random examples from the test set in the [1-5K] frequency bin, another 100 from the [5K-20K] frequency bin and 30 from the [100K-200K] frequency bin, and manually analyzed each of the errors made by all the 4 different variants.

Our analysis first reveals that, in all the cases, about a third of the translations taken as erroneous according to the gold standard are not so in reality. This corresponds to both different morphological variants of the gold standard translations (e.g. *dichiarato/dichiarò*) and other valid translations that were missing in the gold standard (e.g. *climb* → *salita* instead of the gold standard *scalato*). This phenomenon is considerably more pronounced in the first frequency bins, which already have a much higher accuracy according to the gold standard.

As for the actual errors, we observe that nearly a third of them correspond to named entities for all the different variants. Interestingly, the vast majority of the proposed translations in these cases are also named entities (e.g. *Ryan* → *Jason*, *John* → *Paolo*), which are often highly related to the original ones (e.g. *Volvo* → *BMW*, *Olympus* → *Nikon*). While these are clear errors, it is understandable that these methods are unable to discriminate between named entities to this degree based solely on the distributional hypothesis, in particular when it comes to common proper names (e.g. *John*, *Andy*), and one could design alternative strategies to address this issue like taking the edit distance as an additional signal.

For the remaining errors, all systems tend to propose translations that have some degree of relationship with the correct ones, including near-synonyms (e.g. *guidelines* → *raccomandazioni*), antonyms (e.g. *sender* → *destinatario*) and words in the same semantic field (e.g. *nominalism* → *intuizionismo* / *innatismo*, which are all philosophical doctrines). However, there are also a few instances where the relationship is weak or unclear (e.g. *loch* → *giardini*, *sweep* → *serrare*). We also observe a few errors that are related to multiwords or collocations (e.g. *carrier* → *aereo*, presumably related to the multiword *air carrier* / *linea aerea*), as well as some rare word that is repeated across many translations (*Ferruzzi*), which could be attributed to the hubness problem (Dinu et al., 2015; Lazaridou et al., 2015).

All in all, our error analysis reveals that the baseline method of Artetxe et al. (2016) and the proposed algorithm tend to make the same kind of errors regardless of the seed dictionary used by the latter, which reinforces our interpretation in the previous section regarding an underlying optimization objective that is independent from any training dictionary. Moreover, it shows that the quality of the learned mappings is much better than what the raw accuracy numbers might suggest, encouraging the incorporation of these techniques in other applications.

7 Conclusions and future work

In this work, we propose a simple self-learning framework to learn bilingual word embedding mappings in combination with any embedding mapping and dictionary induction technique. Our experiments on bilingual lexicon induction and crosslingual word similarity show that our method is able to learn high quality bilingual embeddings from as little bilingual evidence as a 25 word dictionary or an automatically generated list of numerals, obtaining results that are competitive with state-of-the-art systems using much richer bilingual resources like larger dictionaries or parallel corpora. In spite of its simplicity, a more detailed analysis shows that our method is implicitly optimizing a meaningful objective function that is independent from any bilingual data which, with a better optimization method, might allow to learn bilingual word embeddings in a completely unsupervised manner.

In the future, we would like to delve deeper into this direction and fine-tune our method so it can reliably learn high quality bilingual word embeddings without any bilingual evidence at all. In addition to that, we would like to explore non-linear transformations (Lu et al., 2015) and alternative dictionary induction methods (Dinu et al., 2015; Smith et al., 2017). Finally, we would like to apply our model in the decipherment scenario (Dou et al., 2015).

Acknowledgements

We thank the anonymous reviewers for their insightful comments and Flavio Merenda for his help with the error analysis.

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

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*. Association for Computational Linguistics, Austin, Texas, pages 2289–2294. https://aclweb.org/anthology/D16-1250.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation* 43(3):209–226.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 1–7. http://www.aclweb.org/anthology/P15-2001.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 1818–1827. http://aclweb.org/anthology/C16-1171.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 1853–1861. http://papers.nips.cc/paper/5270-an-autoencoder-approach-to-learning-bilingual-word-representations.pdf.

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*.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying bayesian inference and vector space models for improved decipherment. 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)*. Association for Computational Linguistics, Beijing, China, pages 836–845. http://www.aclweb.org/anthology/P15-1081.

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*. Association for Computational Linguistics, Gothenburg, Sweden, pages 462–471. http://www.aclweb.org/anthology/E14-1049.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*. pages 748–756.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 1459–1474. http://www.aclweb.org/anthology/C12-1089.

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)*. Association for Computational Linguistics, Beijing, China, pages 270–280. http://www.aclweb.org/anthology/P15-1027.

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*. Association for Computational Linguistics, Denver, Colorado, pages 250–256. http://www.aclweb.org/anthology/N15-1028.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Association for Computational Linguistics, Denver, Colorado, pages 151–159. http://www.aclweb.org/anthology/W15-1521.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2016. Adversarial autoencoders. In *Proceedings of the 4rd International Conference on Learning Representations (ICLR 2016), workshop track*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 55–60. http://www.aclweb.org/anthology/P14-5010.

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*. Association for Computational Linguistics, Berlin, Germany, pages 121–126. http://anthology.aclweb.org/W16-1614.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Aditya Mogadala and Achim Rettinger. 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*. Association for Computational Linguistics, San Diego, California, pages 692–702. http://www.aclweb.org/anthology/N16-1083.

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*. Association for Computational Linguistics, Los Angeles, California, pages 921–929. http://www.aclweb.org/anthology/N10-1135.

Samuel L. Smith, David H.P. 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), conference track*.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 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)*. Association for Computational Linguistics, Beijing, China, pages 1713–1722. http://www.aclweb.org/anthology/P15-1165.

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

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 589–598. http://www.aclweb.org/anthology/N16-1072.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1661–1670. http://www.aclweb.org/anthology/P16-1157.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 247–257. http://www.aclweb.org/anthology/P16-1024.

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*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1613–1624. http://www.aclweb.org/anthology/D13-1168.

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

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 119–129. http://www.aclweb.org/anthology/W14-1613.

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*. Association for Computational Linguistics, Denver, Colorado, pages 1006–1011. http://www.aclweb.org/anthology/N15-1104.

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. Association for Computational Linguistics, San Diego, California, pages 1307–1317. http://www.aclweb.org/anthology/N16-1156.

Kai Zhao, Hany Hassan, and Michael Auli. 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*. Association for Computational Linguistics, Denver, Colorado, pages 1527–1536. http://www.aclweb.org/anthology/N15-1176.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1393–1398. http://www.aclweb.org/anthology/D13-1141.