Beyond Offline Mapping:
Learning Cross-lingual Word Embeddings through Context Anchoring

Aitor Ormazabal[1], **Mikel Artetxe**[2], **Aitor Soroa**[1], **Gorka Labaka**[1], **Eneko Agirre**[1]
[1]HiTZ Center, University of the Basque Country (UPV/EHU)
[2]Facebook AI Research
{aitor.ormazabal,a.soroa,gorka.labaka,e.agirre}@ehu.eus
artetxe@fb.com

Abstract

Recent research on cross-lingual word embeddings has been dominated by unsupervised mapping approaches that align monolingual embeddings. Such methods critically rely on those embeddings having a similar structure, but it was recently shown that the separate training in different languages causes departures from this assumption. In this paper, we propose an alternative approach that does not have this limitation, while requiring a weak seed dictionary (e.g., a list of identical words) as the only form of supervision. Rather than aligning two fixed embedding spaces, our method works by fixing the target language embeddings, and learning a new set of embeddings for the source language that are aligned with them. To that end, we use an extension of skip-gram that leverages translated context words as anchor points, and incorporates self-learning and iterative restarts to reduce the dependency on the initial dictionary. Our approach outperforms conventional mapping methods on bilingual lexicon induction, and obtains competitive results in the downstream XNLI task.

1 Introduction

Cross-lingual word embeddings (CLWEs) represent words from two or more languages in a shared space, so that semantically similar words in different languages are close to each other. Early work focused on jointly learning CLWEs in two languages, relying on a strong cross-lingual supervision in the form of parallel corpora (Luong et al., 2015; Gouws et al., 2015) or bilingual dictionaries (Gouws and Søgaard, 2015; Duong et al., 2016). However, these approaches were later superseded by offline mapping methods, which separately train word embeddings in different languages and align them in an unsupervised manner through self-learning (Artetxe et al., 2018; Hoshen and Wolf, 2018) or adversarial training (Zhang et al., 2017; Conneau et al., 2018a).

Despite the advantage of not requiring any parallel resources, mapping methods critically rely on the underlying embeddings having a similar structure, which is known as the *isometry assumption*. Several authors have observed that this assumption does not generally hold, severely hindering the performance of these methods (Søgaard et al., 2018; Nakashole and Flauger, 2018; Patra et al., 2019). In later work, Ormazabal et al. (2019) showed that this issue arises from trying to align separately trained embeddings, as joint learning methods are not susceptible to it.

In this paper, we propose an alternative approach that does not have this limitation, but can still work without any parallel resources. The core idea of our method is to fix the target language embeddings, and learn aligned embeddings for the source language from scratch. This prevents structural mismatches that result from independently training embeddings in different languages, as the learning of the source embeddings is tailored to each particular set of target embeddings. For that purpose, we use an extension of skip-gram that leverages translated context words as anchor points. So as to translate the context words, we start with a weak initial dictionary, which is iteratively improved through self-learning, and we further incorporate a restarting procedure to make our method more robust. Thanks to this, our approach can effectively work without any human-crafted bilingual resources, relying on simple heuristics (automatically generated lists of numerals or identical words) or an existing unsupervised mapping method to build the initial dictionary. Our experiments confirm the effectiveness of our approach, outperforming previous mapping methods on bilingual dictionary induction and obtaining competitive results on zero-shot cross-lingual transfer learning on XNLI.

2 Related work

Word embeddings. Embedding methods learn static word representations based on co-occurrence statistics from a corpus. Most approaches use two different matrices to represent the words and the contexts, which are known as the *input* and *output* vectors, respectively (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017). The output vectors play an auxiliary role, being discarded after training. Our method takes advantage of this fact, leveraging translated output vectors as anchor points to learn cross-lingual embeddings. To that end, we build on the Skip-Gram with Negative Sampling (SGNS) algorithm (Mikolov et al., 2013), which trains a binary classifier to distinguish whether each output word co-occurs with the given input word in the training corpus or was instead sampled from a noise distribution.

Mapping CLWE methods. Offline mapping methods separately train word embeddings for each language, and then learn a mapping to align them into a shared space. Most of these methods align the embeddings through a linear map—often enforcing orthogonality constraints—and, as such, they rely on the assumption that the geometric structure of the separately learned embeddings is similar. This assumption has been shown to fail under unfavorable conditions, severely hindering the performance of these methods (Søgaard et al., 2018; Vulić et al., 2020). Existing attempts to mitigate this issue include learning non-linear maps in a latent space (Mohiuddin et al., 2020), employing maps that are only locally linear (Nakashole, 2018), or learning a separate map for each word (Glavaš and Vulić, 2020). However, all these methods are supervised, and have the same fundamental limitation of aligning a set of separately trained embeddings (Ormazabal et al., 2019).

Self-learning. While early mapping methods relied on a bilingual dictionary to learn the alignment, this requirement was alleviated thanks to self-learning, which iteratively re-induces the dictionary during training. This enabled learning CLWEs in a semi-supervised fashion starting from a weak initial dictionary (Artetxe et al., 2017), or in a completely unsupervised manner when combined with adversarial training (Conneau et al., 2018a) or initialization heuristics (Artetxe et al., 2018; Hoshen and Wolf, 2018). Our proposed method also incorporates a self-learning procedure, showing that this technique can also be effective with non-mapping methods.

Joint CLWE methods. Before the popularization of offline mapping, most CLWE methods extended monolingual embedding algorithms by either incorporating an explicit cross-lingual term in their learning objective, or directly replacing words with their translation equivalents in the training corpus. For that purpose, these methods relied on some form of cross-lingual supervision, ranging from bilingual dictionaries (Gouws and Søgaard, 2015; Duong et al., 2016) to parallel or document-aligned corpora (Luong et al., 2015; Gouws et al., 2015; Vulić and Moens, 2016). More recently, Lample et al. (2018) reported positive results learning regular word embeddings over concatenated monolingual corpora in different languages, relying on identical words as anchor points. Wang et al. (2019) further improved this approach by applying a conventional mapping method afterwards. As shown later in our experiments, our approach outperforms theirs by a large margin.

Freezing. Artetxe et al. (2020) showed that it is possible to transfer an English transformer to a new language by freezing all the inner parameters of the network and learning a new set of embeddings for the new language through masked language modeling. This works because the frozen transformer parameters constrain the resulting representations to be aligned with English. Similarly, our proposed approach uses frozen output vectors in the target language as anchor points to learn aligned embeddings in the source language.

3 Proposed method

Let $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ be the input and output vectors of the ith word in the source language, and $\mathbf{y}_j$ and $\tilde{\mathbf{y}}_j$ be their analogous in the target language.[1] In addition, let D be a bilingual dictionary, where $D(i) = j$ denotes that the ith word in the source language is translated as the jth word in the target language. Our approach first learns the target language embeddings $\{\mathbf{y}_i\}$ and $\{\tilde{\mathbf{y}}_i\}$ monolingually using regular SGNS. Having done that, we learn the source language embeddings $\{\mathbf{x}_i\}$ and $\{\tilde{\mathbf{x}}_i\}$, constraining them to be aligned with the target language embeddings according to the dictionary D. For that purpose, we propose an extension of

[1] Recall that $\{\tilde{\mathbf{x}}_i\}$ and $\{\tilde{\mathbf{y}}_j\}$ are auxiliary, and the goal is to learn aligned $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_j\}$ (see §2).

Algorithm 1 Proposed method

```
Input: $D$ (dictionary), $C_{src}$ (src corpus), $C_{tgt}$ (tgt corpus)
Output: $\{\mathbf{x}_i\}$, $\{\mathbf{y}_i\}$ (aligned src and tgt embs)
Hparams: $T$ (updates), $R$ (restarts), $K$ (re-inductions)
$\{\mathbf{y}_i\}, \{\tilde{\mathbf{y}}_i\} \leftarrow$ SGNS($C_{tgt}$)      ▷ learn target embedings
for $r \leftarrow 1$ to $R$ do                   ▷ iterative restart (§3.3)
    $\{\mathbf{x}_i\}, \{\tilde{\mathbf{x}}_i\} \leftarrow$ RANDOM_INIT()
    for $it \leftarrow 1$ to $T$ do
        $(w_i, w_j) \leftarrow$ NEXT_INSTANCE($C_{src}$)
        BACKPROP($\mathcal{L}(w_i, w_j)$)        ▷ core method (§3.1)
        if $it \mod (T/K) = 0$ then ▷ self-learn (§3.2)
            $D \leftarrow$ REINDUCE($\{\mathbf{x}_i\}$, $\{\mathbf{y}_i\}$)
        end if
    end for
end for
```

SGNS that replaces the output vectors in the source language with their translation equivalents in the target language, which act as anchor points (§3.1). So as to make our method more robust to a weak initial dictionary, we incorporate a self-learning procedure that re-estimates the dictionary during training (§3.2), and perform iterative restarts (§3.3). Algorithm 1 summarizes our method.

3.1 SGNS with cross-lingual anchoring

Given a pair of words (w_i, w_j) co-occurring in the source language corpus, we define a generalized SGNS objective as follows:

$$\mathcal{L}(w_i, w_j) = \log \sigma\left(\mathbf{x}_{w_i} \cdot \mathrm{ctx}(w_j)\right) + \sum_{i=1}^{k} \mathbb{E}_{w_n \sim P_n(w)} \left[\log \sigma\left(-\mathbf{x}_{w_i} \cdot \mathrm{ctx}(w_n)\right)\right]$$

where k is the number of negative samples, $P_n(w)$ is the noise distribution, and $\mathrm{ctx}(w_t)$ is a function that returns the output vector to be used for w_t. In regular SGNS, this function would simply return the output vector of the corresponding word, so that $\mathrm{ctx}(w_t) = \tilde{\mathbf{x}}_{w_t}$. In contrast, our approach replaces it with its counterpart in the target language if w_t is in the dictionary:

$$\mathrm{ctx}(w_t) = \begin{cases} \tilde{\mathbf{y}}_{D(w_t)} & \text{if } w_t \in D \\ \tilde{\mathbf{x}}_{w_t} & \text{otherwise} \end{cases}$$

During training, the replaced vectors $\{\tilde{\mathbf{y}}_i\}$ are kept frozen, acting as anchor points so that the resulting embeddings $\{\mathbf{x}_i\}$ are aligned with their counterparts $\{\mathbf{y}_i\}$ in the target language.

3.2 Self-learning

As shown later in our experiments, the performance of our basic method is largely dependent on the quality of the bilingual dictionary itself. However, this is not different for conventional mapping methods, which also rely on a bilingual dictionary to align separately trained embeddings in different languages. So as to overcome this issue, modern mapping approaches rely on self-learning, which alternates between aligning the embeddings and re-inducing the dictionary in an iterative fashion (Artetxe et al., 2017).

We adopt a similar strategy, and re-induce the dictionary D a total of K times during training, where K is a hyperparameter. To that end, we first obtain the translations for each source word using CSLS retrieval (Conneau et al., 2018a):

$$D(i) = \arg\max_j \mathrm{CSLS}(\mathbf{x}_i, \mathbf{y}_j)$$

Having done that, we discard all entries that do not satisfy the following cyclic consistency condition:[2]

$$i \in D \iff i = \arg\max_k \cos\left(\mathbf{x}_k, \mathbf{y}_{\arg\max_j \cos(\mathbf{x}_i, \mathbf{y}_j)}\right)$$

3.3 Iterative restarts

While self-learning is able to improve a weak initial dictionary throughout training, the method is still susceptible to poor local optima. This can be further exacerbated by the learning rate decay commonly used with SGNS, which makes it increasingly difficult to recover from a poor solution as training progresses. So as to overcome this issue, we sequentially run the entire SGNS training R times, where R is a hyperparameter of the method. We use the output from the previous run as the initial dictionary, but all the other parameters are reset and the full training process is run from scratch.

4 Experimental setup

We next describe the systems explored in our experiments (§4.1), the data and procedure used to train them (§4.2), and the evaluation tasks (§4.3).

4.1 Systems

We compare 3 model families in our experiments:

Offline mapping. This approach learns monolingual embeddings in each of the languages separately, which are then mapped into a common space

[2] We define our cyclic consistency condition over cosine similarity, which we found to be more restrictive than CSLS (in that it discards more entries) and work better in our preliminary experiments.

	en	de	es	fr	fi	ru	zh
Tokens	2,390	860	601	724	91	498	234
Sentences	101	42	22	28	6	25	10

Table 1: Size of the training corpora (millions).

	de-en	es-en	fr-en	fi-en	ru-en	zh-en
Identical	44.8	57.6	63.8	37.7	4.3	3.3
Numerals	1.4	1.6	1.6	2.4	1.1	0.2
Mapping	53.3	67.3	69.7	22.3	28.2	17.1

Table 2: Size of the initial dictionaries (thousands).

through a linear transformation. We experiment with 3 popular methods from the literature: MUSE (Conneau et al., 2018a), ICP (Hoshen and Wolf, 2018) and VecMap (Artetxe et al., 2018). We use the original implementation of each method in their unsupervised mode with default hyperparameters.

Joint learning + offline mapping. This approach jointly learns word embeddings for two languages over their concatenated monolingual corpora, where identical words act as anchor points (Lample et al., 2018). Having done that, the vocabulary is partitioned into one shared and two language specific subsets, which are further aligned through an offline mapping method (Wang et al., 2019). We use the joint_align implementation from the authors with default hyperparameters, which relies on fastText for the joint learning step and MUSE for the mapping step.[3]

Cross-lingual anchoring. Our proposed method, described in Section 3. We explore 3 alternatives to obtain the initial dictionary: **(i) identical words**, where $D_i = j$ if the ith source word and the jth target word are identically spelled, **(ii) numerals**, a subset of the former where identical words are further restricted to be sequences of digits, and **(iii) unsupervised mapping**, where we use the baseline VecMap system described above to induce the initial dictionary.[4] The first two variants make assumptions on the writing system of different languages, which is usually regarded as a weak form of supervision (Artetxe et al., 2017; Søgaard et al., 2018), whereas the latter is strictly unsupervised, yet dependant on an additional system from a different family.

4.2 Data and training details

We learn CLWEs between English and six other languages: German, Spanish, French, Finnish, Russian and Chinese. Following common practice, we use Wikipedia as our training corpus,[5] which we preprocessed using standard Moses scripts, and restrict our vocabulary to the most frequent 200K tokens per language. In the case of Chinese, word segmentation was done using the Stanford Segmenter. Table 1 summarizes the statistics of the resulting corpora, while Table 2 reports the sizes of the initial dictionaries derived from it for our proposed method.

For joint_align, we directly run the official implementation over our tokenized corpus as described above. All the other systems take monolingual embeddings as input, which we learn using the SGNS implementation in word2vec.[6] For our proposed method, we set English as the target language, fix the corresponding monolingual embeddings, and learn aligned embeddings in the source language using our extension of SGNS (§3).[7] We set the number of restarts R to 3, the number of reinductions per restart K to 50, and the number of epochs to $10\frac{\text{\#trg sents}}{\text{\#src sents}}$, which makes sure that the source language gets a similar number of updates to the 10 epochs done for English.[8] For all the other hyperparameters, we use the same values as for the monolingual embeddings. We made all of our development decisions based on preliminary experiments on English-Finnish, without any systematic hyperparameter exploration. Our implementation runs on CPU, except for the dictionary reinduction steps, which run on a single GPU for around one

[3] The original implementation only supports the supervised mode with RCSLS mapping, so we modified it to use MUSE in the unsupervised setting as described in the original paper.

[4] We use CSLS retrieval and apply the cyclic consistency restriction as described in Section 3.2.

[5] We extracted the corpus from the February 2019 dump using the WikiExtractor tool.

[6] We use 10 negative samples, a sub-sampling threshold of 1e-5, 300 dimensions, and 10 epochs. Note that joint_align also learns 300-dimensional vectors, but runs fastText with default hyperparameters under the hood.

[7] In our preliminary experiments, we observed our proposed method to be quite sensitive to which language is the source and which one is the target. We find the language with the largest corpus to perform best as the target, presumably because the corresponding monolingual embeddings are better estimated, so it is more appropriate to fix them and learn aligned embeddings for the other language. Following this observation, we set English as the target language for all pairs, as it is the language with the largest corpus.

[8] For a fair comparison, we also tried using the same number of epochs for the baseline systems, but this performed worse than the reported numbers with 10 epochs.

	de-en		es-en		fr-en		fi-en		ru-en		zh-en		avg
	→	←	→	←	→	←	→	←	→	←	→	←	
OFFLINE MAPPING													
MUSE (Conneau et al., 2018a)	72.9	74.8	83.5	83.0	81.7	82.3	0.3^{*}	0.3^{*}	0.0^{*}	0.3^{*}	39.5	30.9	45.8
ICP (Hoshen and Wolf, 2018)	73.9	75.1	82.5	83.2	80.5	82.3	0.3^{*}	0.3^{*}	59.5	46.1	0.1^{*}	2.8^{*}	48.9
VecMap (Artetxe et al., 2018)	74.5	76.6	83.5	83.3	82.7	83.0	61.9	45.1	**65.7**	49.0	42.4	33.4	65.1
JOINT LEARNING + OFFLINE MAPPING													
Joint_Align (Wang et al., 2019)	70.7	68.7	71.9	69.6	79.2	78.0	33.1	29.1	31.3	25.1	3.6^{*}	18.4	48.2
CROSS-LINGUAL ANCHORING													
Ours (identical init)	76.7	77.9	**86.5**	84.1	**85.0**	84.8	63.3	51.3	65.3	**51.6**	42.1	**38.9**	67.3
Ours (numeral init)	**76.9**	77.7	86.3	84.1	**85.0**	**84.9**	63.6	50.6	64.9	51.4	1.4^{*}	4.9^{*}	61.0
Ours (mapping init)	76.8	**78.1**	86.3	**84.2**	84.9	**84.9**	**64.2**	**51.5**	**65.7**	51.5	**42.5**	38.8	**67.5**

Table 3: Main BLI results on the MUSE dataset (P@1). Asterisks denote divergence ($< 5\%$ P@1).

hour in total.

4.3 Evaluation tasks

As described next, we evaluate our method on two tasks: Bilingual Lexicon Induction (BLI) and Cross-lingual Natural Language Inference (XNLI).

BLI. Following common practice, we induce a bilingual dictionary through CSLS retrieval (Conneau et al., 2018a) for each set of cross-lingual embeddings, and evaluate the precision at 1 (P@1) with respect to the gold standard test dictionary from the MUSE dataset (Conneau et al., 2018a). For the few out-of-vocabulary source words, we revert to copying as a back-off strategy,[9] so our reported numbers are directly comparable to prior work in terms of coverage.

XNLI. We train an English natural language inference model on MultiNLI (Williams et al., 2018), and evaluate the zero-shot cross-lingual transfer performance on the XNLI test set (Conneau et al., 2018b) for the subset of our languages covered by it. To that end, we follow Glavaš et al. (2019) and train an Enhanced Sequential Inference Model (ESIM) on top of our original English embeddings, which are kept frozen during training. At test time, we transfer into the rest of the languages by plugging in the corresponding aligned embeddings. Note that we use the exact same English model for our proposed method and the baseline MUSE and ICP systems,[10] which only differ in the set of aligned embeddings used for cross-lingual transfer. In contrast, VecMap and joint_align also manipulate the target English embeddings, which would require training a separate model for each language pair, so we decide to exclude them from this set of experiments.[11]

5 Results

We next discuss our main results on BLI (§5.1) and XNLI (§5.2), followed by our ablation study (§5.3) and error analysis (§5.4) on BLI.

5.1 BLI

Table 3 comprises our main BLI results. We observe that our method obtains the best results in all directions (matched by VecMap in Russian-English), outperforming the strongest baseline by 2.4 points on average for the mapping based initialization. Our improvements are more pronounced in the backward direction (3.1 points on average) but still substantial in the forward direction (1.7 points on average).

It is worth noting that some systems fail to converge to a good solution for the most challenging language pairs. This includes our proposed method in the case of Chinese-English when using the numeral-based initialization, which we attribute to the smaller size of the initial dictionary (only 244 entries, see Table 2). Other than that, we observe that our approach obtains very similar results regardless of the initial dictionary. Quite remarkably,

[9] This has a negligible impact in practice, as it accounts for less than 1.4% of the test cases. Moreover, all of our systems use the same underlying vocabulary, so they are affected in the exact same way.

[10] This is possible because they all fix the target language embeddings (English in this case) and align the embeddings in the source language with them, either through mapping (MUSE, ICP) or learning from scratch (ours).

[11] In addition to the computational overhead, having separate models introduces some variance, while our comparison is more direct.

	de-en		es-en		fr-en		ru-en		avg
	→	←	→	←	→	←	→	←	
Conneau et al. (2018a)	72.2	74.0	83.3	81.7	82.1	82.3	59.1	44.0	72.3
Hoshen and Wolf (2018)	73.0	74.7	84.1	82.1	82.9	82.3	61.8	47.5	73.6
Grave et al. (2018)	73.3	75.4	84.1	82.8	82.9	82.6	59.1	43.7	73.0
Alvarez-Melis and Jaakkola (2018)	72.8	71.9	80.4	81.7	78.9	81.3	43.7	45.1	69.5
Yang et al. (2018)	70.3	71.5	79.3	79.9	78.9	78.4	-	-	-
Mukherjee et al. (2018)	-	-	79.2	**84.5**	-	-	-	-	-
Alvarez-Melis et al. (2018)	71.1	73.8	81.8	81.3	81.6	82.9	55.4	41.7	71.2
Xu et al. (2018)	67.0	69.3	77.8	79.5	75.5	77.9	-	-	-
Wang et al. (2019)	72.2	74.2	84.2	81.4	83.6	82.8	58.3	45.0	72.7
Zhou et al. (2019)	74.4	77.2	84.9	82.8	83.5	83.1	63.6	49.2	74.8
Li et al. (2020)	74.3	75.3	84.6	82.4	83.7	82.6	-	-	-
Ours (mapping init)	**76.8**	**78.1**	**86.3**	84.2	**84.9**	**84.9**	**65.7**	**51.5**	**76.6**

Table 4: BLI results on MUSE dataset in comparison with prior published results (P@1). All systems are fully unsupervised (except that of Zhou et al. (2019), which uses identical words as a seed dictionary), and use SGNS embeddings trained on Wikipedia.

	en	de	es	fr	ru	zh
MUSE	**73.9**	65.0	68.1	**67.9**	39.1*	**61.4**
ICP	**73.9**	62.2	64.2	65.7	59.4	36.1*
Ours (identical init)	**73.9**	65.0	**68.7**	67.1	**63.5**	49.8
Ours (numeral init)	**73.9**	65.0	68.6	67.1	63.3	34.9*
Ours (mapping init)	**73.9**	**65.1**	68.6	67.0	**63.5**	49.4

Table 5: XNLI results (accuracy). Asterisks denote divergence ($< 5\%$ P@1 in BLI).

Basic method (identical init)	53.9
+ *self-learning*	66.9
+ *iterative restarts*	67.3
Basic method (numeral init)	2.6
+ *self-learning*	53.9
+ *iterative restarts*	61.0
Basic method (mapping init)	67.5
+ *self-learning*	67.5
+ *iterative restarts*	67.5

Table 6: Ablation results on BLI (average P@1)

the variant using VecMap for initialization (*mapping init*) is substantially stronger than VecMap itself despite not using any additional training signal.

So as to put our results into perspective, Table 4 compares them to previous numbers reported in the literature. Note that the numbers are comparable in terms of coverage and all systems use Wikipedia as the training corpus, although they might differ on the specific dump used and the preprocessing details.[12] As it can be seen, our approach obtains the best results by a substantial margin.[13]

5.2 XNLI

We report our XNLI results in Table 5. We observe that our method is competitive with the baseline mapping systems, achieving the best results on 3 out of the 5 transfer languages by a small margin. Nevertheless, it significantly lags behind MUSE on Chinese, even if the exact same set of cross-lingual embeddings performed better than MUSE at BLI. While striking, similar discrepancies between BLI and XNLI performance where also observed in previous studies (Glavaš et al., 2019). Finally, we observe that the initial dictionary has a negligible impact in the performance of our proposed method, which supports the idea that our approach converges to a similar solution given any reasonable initialization.

5.3 Ablation study

So as to understand the role of self-learning and the iterative restarts in our approach, we perform an ablation study and report our results in Table 6. We observe that the contribution of these components is greatly dependant on the initial dictionary. For the numeral initialization, the basic method works poorly, and both extensions bring large improvements. In contrast, the identical initialization

[12] In particular, most mapping methods use the official Wikipedia embeddings from fastText. Unfortunately, the preprocessed corpus used to train these embeddings is not public, so works that explore other approaches, like ours, need to use their own pre-processed copy of Wikipedia.

[13] Artetxe et al. (2019) report even stronger results based on unsupervised machine translation instead of direct retrieval with CLWEs. Note, however, that their method still relies on cross-lingual embeddings to build the underlying phrase-table, so our improvements should be largely orthogonal to theirs.

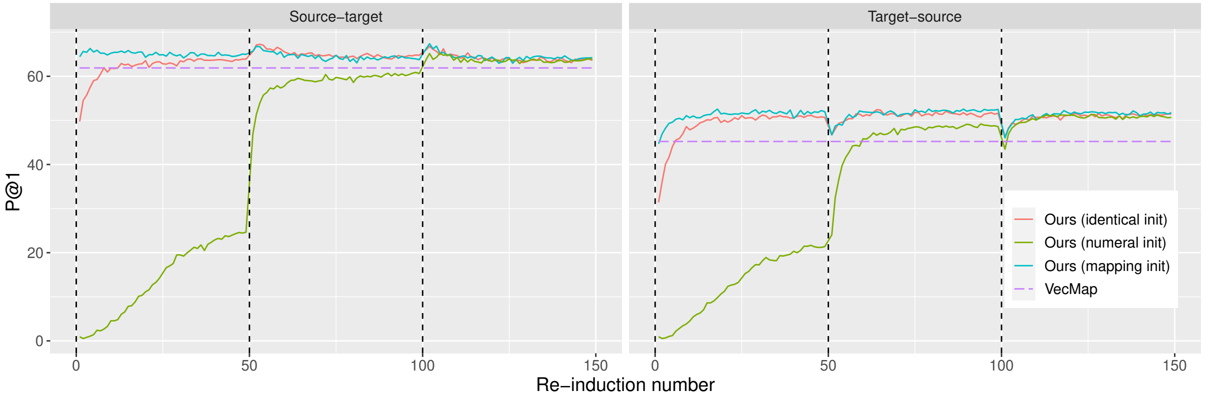

Figure 1: Finnish-English learning curves (BLI P@1). The iterative restarts happen at the vertical lines.

does not benefit from iterative restarts, but self-learning still plays a major role. In the case of the mapping-based initialization, the basic method is already very competitive. This suggests that both the self-learning and the iterative restarts are helpful to make the method more robust to a weak initialization, and have a minor impact otherwise.

In order to better understand the underlying learning dynamics, we analyze the learning curves for Finnish-English in Figure 1. We observe that, when the initial dictionary is strong, our method surpasses the baseline and stabilizes early. In contrast, convergence is much slower when using the weak numeral-based initialization, and the iterative restarts are critical to escape poor local optima.

5.4 Error analysis

So as to better understand where our improvements in BLI are coming from, we perform an error analysis on the Spanish-English direction. To that end, we manually inspect the 69 instances for which our method (with mapping-based initialization) produced a correct translation while VecMap failed according to the gold standard, as well as the 26 instances for which the opposite was true. We then categorize these errors into several types, which are summarized in Table 7.

We observe that, in 52.6% of the 95 analyzed instances, the translation produced by our method is identical to the source word, while this percentage goes down to 4.2% for VecMap. This tendency of our approach to copy its input is striking, as the model has no notion about the words being identically spelled.[14] A large portion of these cases correspond to named entities, where copying is the right behavior, while VecMap outputs a different proper noun. There are also some instances where the input word is in the target language,[15] which can be considered an artifact of the dataset, but copying also seems the most reasonable behavior in these cases. Finally, there are also a few cases where the input word is present in the target vocabulary, which is selected by our method and counted as an error. Once again, we consider these to be an artifact of the dataset, as copying seems a reasonable choice if the input word is considered to be part of the target language vocabulary. The remaining cases where neither method copies mostly correspond to common errors, where one of the systems (most often VecMap) outputs a semantically related but incorrect translation. However, there are also a few instances where both translations are correct, but one of them is missing in the gold standard.

With the aim to understand the impact of identical words in our original results, we re-evaluated the systems using a filtered version of the MUSE gold standard dictionaries, where we removed all source words that were included in the set of candidate translations. In order to be fair, we filtered out identical words from the output of the system, reverting to the second highest-ranked translation whenever the first one is identical to the source word. The results for the strongest system in each family are shown in Table 8. Even if the margin of improvement is reduced compared to Table 3, the best results are still obtained by our proposed method, bringing an average improvement

[14] The variants of our system with identical or numeral initialization do indirectly see this signal, but the one analyzed here is initialized with the VecMap mapping.

[15] English words will often appear in other languages as part of named entities (e.g., “pink” as part of “Pink Floyd”), which explains the presence of such words in the Spanish vocabulary.

Gold standard	Type	Cases	Examples		
			Source	VecMap	Ours
Ours right – VecMap wrong	Common errors	30.5%	derrotas campeona	victories medalist	defeats champion
	Named entity, ours copies	21.1%	philadelphia susana	pittsburgh beatriz	philadelphia susana
	EN word in ES vocab, ours copies	15.8%	pink space	tangerine sci	pink space
	Gap in gold standard	5.3%	adecuada marquesa	appropriate marchioness	adequate marquise
VecMap right – Ours wrong	Common errors	15.8%	conservadores noveno	conservatives ninth	liberals tenth
	ES word in EN vocab, ours copies	7.4%	calzada cantera	roadway quarry	calzada cantera
	Gap in gold standard	4.2%	ferroviario situados	railway situated	rail positioned

Table 7: BLI error analysis on Spanish-English. See Section 5.4 for details.

	de-en		es-en		fr-en		fi-en		ru-en		zh-en		avg
	→	←	→	←	→	←	→	←	→	←	→	←	
VecMap (Artetxe et al., 2018)	68.3	70.2	85.1	79.4	80.8	78.1	**58.4**	38.9	**66.1**	48.6	45.0	34.5	62.8
Joint_Align (Wang et al., 2019)	57.0	53.3	63.0	57.4	70.2	64.4	4.0^*	0.7^*	31.3	22.4	3.5^*	0.9^*	35.7
Ours (identical init)	68.9	72.2	**86.0**	80.7	81.5	80.0	54.0	41.0	65.7	50.9	44.6	**38.1**	63.6
Ours (mapping init)	**68.9**	**72.3**	85.8	**80.8**	**81.4**	**80.2**	55.4	**41.6**	**66.1**	**51.0**	**45.1**	37.9	**63.9**

Table 8: BLI results on MUSE with identical words removed (P@1). Asterisks denote divergence ($< 5\%$ P@1).

of 1.1 points. It is also worth noting that joint_align, which shares a portion of the vocabulary for both languages (and will thus translate all words in the shared vocabulary identically), suffers a large drop in performance. This highlights the importance of accompanying quantitative BLI evaluation with an error analysis as urged by previous studies (Kementchedjhieva et al., 2019).

6 Conclusions and future work

Our approach for learning CLWEs addresses the main limitations of both offline mapping and joint learning methods. Different from mapping approaches, it does not suffer from structural mismatches arising from independently training embeddings in different languages, as it works by constraining the learning of the source embeddings so they are aligned with the target ones. At the same time, unlike previous joint methods, our system can work without any parallel resources, relying on numerals, identical words or an existing mapping method for the initialization. We achieve this by combining cross-lingual anchoring with self-learning and iterative restarts. While recent research on CLWEs has been dominated by mapping approaches, our work shows that the fundamental techniques that popularized these methods (e.g., the use of self-learning to relax the need for cross-lingual supervision) can also be effective beyond this paradigm.

Despite its simplicity, our experiments on BLI show the superiority of our method when compared to previous mapping systems. We complement these results with additional experiments on a downstream task, where our method obtains competitive results, as well as an ablation study and a systematic error analysis. We identify a striking tendency of our method to translate words identically, even if it has no notion of the words being identically spelled. Thanks to this, our method is particularly strong at translating named entities, but we show that our improvements are not limited to this phenomenon. These insights confirm the value of accompanying quantitative results on BLI with qualitative evaluation (Kementchedjhieva et al., 2019) and/or other tasks (Glavaš et al., 2019).

In the future, we would like to further explore CLWE methods that go beyond the currently dominant mapping paradigm. In particular, we would like to remove the requirement of a seed dictionary altogether by using adversarial learning, and explore more elaborated context translation and dictionary re-induction schemes.

Acknowledgments

Aitor Ormazabal, Aitor Soroa, Gorka Labaka and Eneko Agirre were supported by the Basque Government (excellence research group IT1343-19 and DeepText project KK-2020/00088), project BigKnowledge (*Ayudas Fundación BBVA a equipos de investigación científica 2018*) and the Spanish MINECO (project DOMINO PGC2018-102041-B-I00 MCIU/AEI/FEDER, UE). Aitor Ormazabal was supported by a doctoral grant from the Spanish MECD.

References

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.

David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. 2018. Towards optimal transport with global invariances. *arXiv preprint arXiv:1806.09277*.

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. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555, Online. Association for Computational Linguistics.

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.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado. Association for Computational Linguistics.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with wasserstein procrustes. *arXiv preprint arXiv:1805.11222*.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium. Association for Computational Linguistics.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3336–3341, Hong Kong, China. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Yanyang Li, Yingfeng Luo, Ye Lin, Quan Du, Huizhen Wang, Shujian Huang, Tong Xiao, and Jingbo Zhu. 2020. A simple and effective approach to robust unsupervised bilingual dictionary induction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5990–6001, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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*, pages 151–159. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. 2020. LNMap: Departures from isomorphic assumption in bilingual lexicon induction through nonlinear mapping in latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2712–2723, Online. Association for Computational Linguistics.

Tanmoy Mukherjee, Makoto Yamada, and Timothy Hospedales. 2018. Learning unsupervised word translations without adversaries. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 627–632, Brussels, Belgium. Association for Computational Linguistics.

Ndapa Nakashole. 2018. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522, Brussels, Belgium. Association for Computational Linguistics.

Ndapa Nakashole and Raphael Flauger. 2018. Characterizing departures from linearity in word translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 221–227, Melbourne, Australia. Association for Computational Linguistics.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. 2019. BLISS in non-isometric embedding spaces.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788. Association for Computational Linguistics.

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.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2019. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. *arXiv preprint arXiv:1910.04708*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng Yang, Fuli Luo, Shuangzhi Wu, Jingjing Xu, Dongdong Zhang, and Xu Sun. 2018. Learning unsupervised word mapping by maximizing mean discrepancy. *arXiv preprint arXiv:1811.00275*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. 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.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density matching for bilingual word embedding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1588–1598, Minneapolis, Minnesota. Association for Computational Linguistics.