# Unsupervised Multilingual Word Embeddings

**Xilun Chen**
Department of Computer Science
Cornell Unversity
Ithaca, NY, 14853, USA
xlchen@cs.cornell.edu

**Claire Cardie**
Department of Computer Science
Cornell Unversity
Ithaca, NY, 14853, USA
cardie@cs.cornell.edu

## Abstract

Multilingual Word Embeddings (MWEs) represent words from multiple languages in a single distributional vector space. Unsupervised MWE (UMWE) methods acquire multilingual embeddings without cross-lingual supervision, which is a significant advantage over traditional supervised approaches and opens many new possibilities for low-resource languages. Prior art for learning UMWEs, however, merely relies on a number of independently trained Unsupervised Bilingual Word Embeddings (UBWEs) to obtain multilingual embeddings. These methods fail to leverage the interdependencies that exist among many languages. To address this shortcoming, we propose a fully unsupervised framework for learning MWEs[1] that directly exploits the relations between all language pairs. Our model substantially outperforms previous approaches in the experiments on multilingual word translation and cross-lingual word similarity. In addition, our model even beats supervised approaches trained with cross-lingual resources.

## 1 Introduction

Continuous distributional word representations (Turian et al., 2010) have become a common technique across a wide variety of NLP tasks. Recent research, moreover, proposes cross-lingual word representations (Klementiev et al., 2012; Mikolov et al., 2013a) that create a shared embedding space for words across *two* (Bilingual Word Embeddings, BWE) or *more* languages (Multilingual Word Embeddings, MWE). Words from different languages with similar meanings will be close to one another in this cross-lingual embedding space. These embeddings have been found beneficial for a number of cross-lingual and even monolingual NLP tasks (Faruqui and Dyer, 2014; Ammar et al., 2016).

[1]Code: https://github.com/ccsasuke/umwe

The most common form of cross-lingual word representations is the BWE, which connects the lexical semantics of two languages. Traditionally for training BWEs, cross-lingual supervision is required, either in the form of parallel corpora (Klementiev et al., 2012; Zou et al., 2013), or in the form of bilingual lexica (Mikolov et al., 2013a; Xing et al., 2015). This makes learning BWEs for low-resource language pairs much more difficult. Fortunately, there are attempts to reduce the dependence on bilingual supervision by requiring a very small parallel lexicon such as identical character strings (Smith et al., 2017), or numerals (Artetxe et al., 2017). Furthermore, recent work proposes approaches to obtain unsupervised BWEs without relying on *any* bilingual resources (Zhang et al., 2017; Lample et al., 2018b).

In contrast to BWEs that only focus on a pair of languages, MWEs instead strive to leverage the interdependencies among multiple languages to learn a multilingual embedding space. MWEs are desirable when dealing with multiple languages simultaneously and have also been shown to improve the performance on some bilingual tasks thanks to its ability to acquire knowledge from other languages (Ammar et al., 2016; Duong et al., 2017). Similar to training BWEs, cross-lingual supervision is typically needed for training MWEs, and the prior art for obtaining *fully unsupervised* MWEs simply maps all the languages independently to the embedding space of a chosen target language[2] (usually English) (Lample et al., 2018b). There are downsides, however, when using a single fixed target language with no interaction between any of the two source languages. For instance, French and Italian are very similar, and the fact that each of them is individually converted to a less similar language, English for example, in

[2]Henceforth, we refer to this method as BWE-Pivot as the target language serves as a pivot to connect other languages.



*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270
Brussels, Belgium, October 31 - November 4, 2018. 

order to produce a shared embedding space will inevitably degrade the quality of the MWEs.

For certain multilingual tasks such as translating between any pair of $N$ given languages, another option for obtaining UMWEs exists. One can directly train UBWEs for each of such language pairs (referred to as BWE-Direct). This is seldom used in practice, since it requires training $O(N^2)$ BWE models as opposed to only $O(N)$ in BWE-Pivot, and is too expensive for most use cases. Moreover, this method still does not fully exploit the language interdependence. For example, when learning embeddings between French and Italian, BWE-Direct only utilizes information from the pair itself, but other Romance languages such as Spanish may also provide valuable information that could improve performance.

In this work, we propose a novel unsupervised algorithm to train MWEs using *only* monolingual corpora (or equivalently, monolingual word embeddings). Our method exploits the interdependencies between any two languages and maps all monolingual embeddings into a shared multilingual embedding space via a two-stage algorithm consisting of (i) Multilingual Adversarial Training (`MAT`) and (ii) Multilingual Pseudo-Supervised Refinement (`MPSR`). As shown by experimental results on multilingual word translation and cross-lingual word similarity, our model is as efficient as BWE-Pivot yet outperforms both BWE-Pivot and BWE-Direct despite the latter being much more expensive. In addition, our model achieves a higher overall performance than state-of-the-art *supervised* methods in these experiments.

## 2 Related Work

There is a plethora of literature on learning cross-lingual word representations, focusing either on a pair of languages, or multiple languages at the same time (Klementiev et al., 2012; Zou et al., 2013; Mikolov et al., 2013a; Gouws et al., 2015; Coulmance et al., 2015; Ammar et al., 2016; Duong et al., 2017, *inter alia*). One shortcoming of these methods is the dependence on cross-lingual supervision such as parallel corpora or bilingual lexica. Abundant research efforts have been made to alleviate such dependence (Vulić and Moens, 2015; Artetxe et al., 2017; Smith et al., 2017), but consider only the case of a single pair of languages (BWEs). Furthermore, fully unsupervised methods exist for learning BWEs (Zhang et al., 2017; Lample et al., 2018b; Artetxe et al., 2018a). For unsupervised MWEs, however, previous methods merely rely on a number of independent BWEs to separately map each language into the embedding space of a chosen target language (Smith et al., 2017; Lample et al., 2018b).

Adversarial Neural Networks have been successfully applied to various cross-lingual NLP tasks where annotated data is not available, such as cross-lingual text classification (Chen et al., 2016), unsupervised BWE induction (Zhang et al., 2017; Lample et al., 2018b) and unsupervised machine translation (Lample et al., 2018a; Artetxe et al., 2018b). These works, however, only consider the case of two languages, and our `MAT` method (§3.1) is a generalization to multiple languages.

Mikolov et al. (2013a) first propose to learn cross-lingual word representations by learning a linear mapping between the monolingual embedding spaces of a pair of languages. It has then been observed that enforcing the linear mapping to be orthogonal could significantly improve performance (Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017). These methods solve a linear equation called the orthogonal Procrustes problem for the optimal orthogonal linear mapping between two languages, given a set of word pairs as supervision. Artetxe et al. (2017) find that when using weak supervision (e.g. digits in both languages), applying this Procrustes process iteratively achieves higher performance. Lample et al. (2018b) adopt the iterative Procrustes method with pseudo-supervision in a fully unsupervised setting and also obtain good results. In the MWE task, however, the multilingual mappings no longer have a closed-form solution, and we hence propose the `MPSR` algorithm (§3.2) for learning multilingual embeddings using gradient-based optimization methods.

## 3 Model

In this work, our goal is to learn a single multilingual embedding space for $N$ languages, without relying on *any* cross-lingual supervision. We assume that we have access to monolingual embeddings for each of the $N$ languages, which can be obtained using unlabeled monolingual corpora (Mikolov et al., 2013b; Bojanowski et al., 2017). We now present our unsupervised MWE (UMWE) model that jointly maps the monolingual embeddings of all $N$ languages into a single

space by explicitly leveraging the interdependencies between arbitrary language pairs, but is computationally as efficient as learning $O(N)$ BWEs (instead of $O(N^2)$).

Denote the set of languages as $\mathscr{L}$ with $|\mathscr{L}| = N$. Suppose for each language $l \in \mathscr{L}$ with vocabulary $\mathcal{V}_l$, we have a set of $d$-dimensional monolingual word embeddings $\mathcal{E}_l$ of size $|\mathcal{V}_l| \times d$. Let $\mathcal{S}_l$ denote the monolingual embedding space for $l$, namely the distribution of the monolingual embeddings of $l$. If a set of embeddings $\mathcal{E}$ are in an embedding space $\mathcal{S}$, we write $\mathcal{E} \vdash \mathcal{S}$ (e.g. $\forall l : \mathcal{E}_l \vdash \mathcal{S}_l$). Our models learns a set of encoders $\mathcal{M}_l$, one for each language $l$, and the corresponding decoders $\mathcal{M}_l^{-1}$. The encoders map all $\mathcal{E}_l$ to a single target space $\mathcal{T}$: $\mathcal{M}_l(\mathcal{E}_l) \vdash \mathcal{T}$. On the other hand, a decoder $\mathcal{M}_l^{-1}$ maps an embedding in $\mathcal{T}$ back to $\mathcal{S}_l$.

Previous research (Mikolov et al., 2013a) shows that there is a strong linear correlation between the vector spaces of two languages, and that learning a complex non-linear neural mapping does not yield better results. Xing et al. (2015) further show that enforcing the linear mappings to be orthogonal matrices achieves higher performance. Therefore, we let our encoders $\mathcal{M}_l$ be orthogonal linear matrices, and the corresponding decoders can be obtained by simply taking the transpose: $\mathcal{M}_l^{-1} = \mathcal{M}_l^\top$. Thus, applying the encoder or decoder to an embedding vector is accomplished by multiplying the vector with the encoder/decoder matrix.

Another benefit of using linear encoders and decoders (also referred to as *mappings*) is that we can learn $N - 1$ mappings instead of $N$ by choosing the target space $\mathcal{T}$ to be the embedding space of a specific language (denoted as the *target language*) without losing any expressiveness of the model. Given a MWE with an arbitrary $\mathcal{T}$, we can construct an equivalent one with only $N-1$ mappings by multiplying the encoders of each language $\mathcal{M}_l$ to the decoder of the chosen target language $\mathcal{M}_t^\top$:

$$\mathcal{M}'_t = \mathcal{M}_t^\top \mathcal{M}_t = I$$
$$\mathcal{M}'_l \mathcal{E}_l = (\mathcal{M}_t^\top \mathcal{M}_l)\mathcal{E}_l \vdash \mathcal{S}_t$$

where $I$ is the identity matrix. The new MWE is isomorphic to the original one.

We now present the two major components of our approach, Multilingual Adversarial Training (§3.1) and Multilingual Pseudo-Supervised Refinement (§3.2).

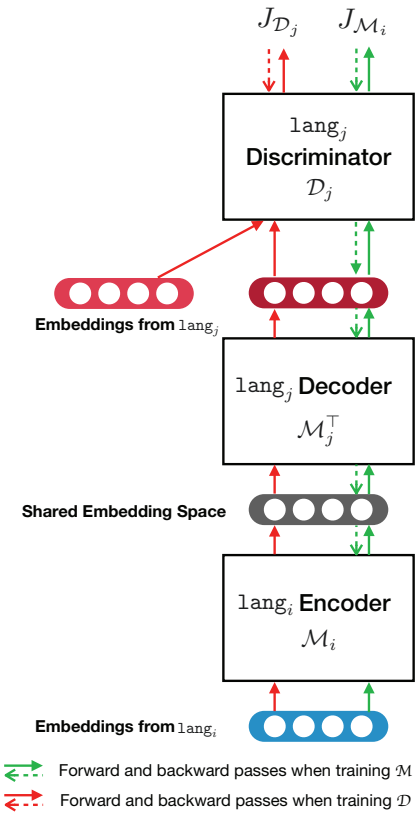


Figure 1: Multilingual Adversarial Training (Algorithm 1). lang$_i$ and lang$_j$ are two randomly selected languages at each training step. $J_{\mathcal{D}_j}$ and $J_{\mathcal{M}_i}$ are the objectives of $\mathcal{D}_j$ and $\mathcal{M}_i$, respectively (Eqn. 1 and 2).

### 3.1 Multilingual Adversarial Training

In this section, we introduce an adversarial training approach for learning multilingual embeddings without cross-lingual supervision. Adversarial Training is a powerful technique for minimizing the divergence between complex distributions that are otherwise difficult to directly model (Goodfellow et al., 2014). In the cross-lingual setting, it has been successfully applied to unsupervised cross-lingual text classification (Chen et al., 2016) and unsupervised bilingual word embedding learning (Zhang et al., 2017; Lample et al., 2018b). However, these methods only consider one pair of languages at a time, and do not fully exploit the cross-lingual relations in the multilingual setting.

Figure 1 shows our Multilingual Adversarial Training (MAT) model and the training procedure is described in Algorithm 1. Note that as explained in §3, the encoders and decoders adopted in practice are orthogonal linear mappings while the shared embedding space is chosen to be the same space as a selected target language.

**Algorithm 1** Multilingual Adversarial Training

**Require:** Vocabulary $\mathcal{V}_i$ for each language $\texttt{lang}_i \in \mathscr{L}$. Hyperparameter $k \in \mathbb{N}$.

1: **repeat**
2: ▷ $\mathcal{D}$ iterations
3: **for** $\texttt{diter} = 1$ to $k$ **do**
4: $\texttt{loss}_d = 0$
5: **for all** $\texttt{lang}_j \in \mathscr{L}$ **do**
6: Select at random $\texttt{lang}_i \in \mathscr{L}$
7: Sample a batch of words $x_i \sim \mathcal{V}_i$
8: Sample a batch of words $x_j \sim \mathcal{V}_j$
9: $\hat{x}_t = \mathcal{M}_i(x_i)$ ▷ encode to $\mathcal{T}$
10: $\hat{x}_j = \mathcal{M}_j^\top(\hat{x}_t)$ ▷ decode to $\mathcal{S}_j$
11: $y_j = \mathcal{D}_j(x_j)$ ▷ real vectors
12: $\hat{y}_j = \mathcal{D}_j(\hat{x}_j)$ ▷ converted vectors
13: $\texttt{loss}_d \mathrel{+}= L_d(1, y_j) + L_d(0, \hat{y}_j)$
14: Update all $\mathcal{D}$ parameters to minimize $\texttt{loss}_d$
15: ▷ $\mathcal{M}$ iteration
16: $\texttt{loss} = 0$
17: **for all** $\texttt{lang}_i \in \mathscr{L}$ **do**
18: Select at random $\texttt{lang}_j \in \mathscr{L}$
19: Sample a batch of words $x_i \sim \mathcal{V}_i$
20: $\hat{x}_t = \mathcal{M}_i(x_i)$ ▷ encode to $\mathcal{T}$
21: $\hat{x}_j = \mathcal{M}_j^\top(\hat{x}_t)$ ▷ decode to $\mathcal{S}_j$
22: $\hat{y}_j = \mathcal{D}_j(\hat{x}_j)$
23: $\texttt{loss} \mathrel{+}= L_d(1, \hat{y}_j)$
24: Update all $\mathcal{M}$ parameters to minimize $\texttt{loss}$
25: $\texttt{orthogonalize}(\mathcal{M})$ ▷ see §3.3
26: **until** convergence

In order to learn a multilingual embedding space without supervision, we employ a series of *language discriminators* $\mathcal{D}_l$, one for each language $l \in \mathscr{L}$. Each $\mathcal{D}_l$ is a binary classifier with a sigmoid layer on top, and is trained to identify how likely a given vector is from $\mathcal{S}_l$, the embedding space of language $l$. On the other hand, to train the mappings, we convert a vector from a random language $\texttt{lang}_i$ to another random language $\texttt{lang}_j$ (via the target space $\mathcal{T}$ first). The objective of the mappings is to confuse $\mathcal{D}_j$, the language discriminator for $\texttt{lang}_j$, so the mappings are updated in a way that $\mathcal{D}_j$ cannot differentiate the converted vectors from the real vectors in $\mathcal{S}_j$. This multilingual objective enables us to explicitly exploit the relations between all language pairs during training, leading to improved performance.

Formally, for any language $\texttt{lang}_j$, the objective that $\mathcal{D}_j$ is minimizing is:

$$J_{\mathcal{D}_j} = \mathop{\mathbb{E}}_{i \sim \mathscr{L}} \mathop{\mathbb{E}}_{\substack{x_i \sim \mathcal{S}_i \\ x_j \sim \mathcal{S}_j}} \Big[\!\Big[ L_d\left(1, \mathcal{D}_j(x_j)\right) + L_d\left(0, \mathcal{D}_j(\mathcal{M}_j^\top \mathcal{M}_i x_i)\right) \Big]\!\Big] \quad (1)$$

where $L_d(y, \hat{y})$ is the loss function of $\mathcal{D}$, which is chosen as the *cross entropy loss* in practice. $y$ is the language label with $y = 1$ indicates a real embedding from that language.

Furthermore, the objective of $\mathcal{M}_i$ for $\texttt{lang}_i$ is:

$$J_{\mathcal{M}_i} = \mathop{\mathbb{E}}_{j \sim \mathscr{L}} \mathop{\mathbb{E}}_{\substack{x_i \sim \mathcal{S}_i \\ x_j \sim \mathcal{S}_j}} L_d\left(1, \mathcal{D}_j(\mathcal{M}_j^\top \mathcal{M}_i x_i)\right) \quad (2)$$

where $\mathcal{M}_i$ strives to make $\mathcal{D}_j$ believe that a converted vector to $\texttt{lang}_j$ is instead real. This adversarial relation between $\mathcal{M}$ and $\mathcal{D}$ stimulates $\mathcal{M}$ to learn a shared multilingual embedding space by making the converted vectors look as authentic as possible so that $\mathcal{D}$ cannot predict whether a vector is a genuine embedding from a certain language or converted from another language via $\mathcal{M}$.

In addition, we allow $\texttt{lang}_i$ and $\texttt{lang}_j$ to be the same language in (1) and (2). In this case, we are encoding a language to $\mathcal{T}$ and back to itself, essentially forming an adversarial autoencoder (Makhzani et al., 2015), which is reported to improve the model performance (Zhang et al., 2017). Finally, on Line 5 and 17 in Algorithm 1, a for loop is used instead of random sampling. This is to ensure that in each step, every discriminator (or mapping) is getting updated at least once, so that we do not need to increase the number of training iterations when adding more languages. Computationally, when compared to the BWE-Pivot and BWE-Direct baselines, one step of MAT training costs similarly to $N$ BWE training steps, and in practice we train MAT for the same number of iterations as training the baselines. Therefore, MAT training scales linearly with the number of languages similar to BWE-Pivot (instead of quadratically as in BWE-Direct).

### 3.2 Multilingual Pseudo-Supervised Refinement

Using MAT, we are able to obtain UMWEs with reasonable quality, but they do not yet achieve state-of-the-art performance. Previous research on learning unsupervised BWEs (Lample et al., 2018b) observes that the embeddings obtained from adversarial training do a good job aligning the frequent words between two languages, but performance degrades when considering the full vocabulary. They hence propose to use an iterative refinement method (Artetxe et al., 2017) to repeatedly refine the embeddings obtained from the adversarial training. The idea is that we can anchor on the more accurately predicted relations between frequent words to improve the mappings learned by adversarial training.

**Algorithm 2** Multilingual Pseudo-Supervised Refinement

**Require:** A set of (pseudo-)supervised lexica of word pairs between each pair of languages $\texttt{Lex}(\texttt{lang}_i, \texttt{lang}_j)$.

**repeat**
  $\texttt{loss} = 0$
  **for all** $\texttt{lang}_i \in \mathscr{L}$ **do**
    Select at random $\texttt{lang}_j \in \mathscr{L}$
    Sample $(x_i, x_j) \sim \texttt{Lex}(\texttt{lang}_i, \texttt{lang}_j)$
    $t_i = \mathcal{M}_i(x_i)$ ▷ encode $x_i$
    $t_j = \mathcal{M}_j(x_j)$ ▷ encode $x_j$
    $\texttt{loss} \mathrel{+}= L_r(t_i, t_j)$ ▷ refinement loss
  Update all $\mathcal{M}$ parameters to minimize $\texttt{loss}$
  $\texttt{orthogonalize}(\mathcal{M})$ ▷ see §3.3
**until** convergence

When learning MWEs, however, it is desirable to go beyond aligning each language with the target space individually, and instead utilize the relations between all languages as we did in MAT. Therefore, we in this section propose a generalization of the existing refinement methods to incorporate a multilingual objective.

In particular, MAT can produce an approximately aligned embedding space. As mentioned earlier, however, the training signals from $\mathcal{D}$ for rare words are noisier and may lead to worse performance. Thus, the idea of Multilingual Pseudo-Supervised Refinement (MPSR) is to induce a dictionary of highly confident word pairs for every language pair, used as pseudo supervision to improve the embeddings learned by MAT. For a specific language pair $(\texttt{lang}_i, \texttt{lang}_j)$, the pseudo-supervised lexicon $\texttt{Lex}(\texttt{lang}_i, \texttt{lang}_j)$ is constructed from *mutual nearest neighbors* between $\mathcal{M}_i\mathcal{E}_i$ and $\mathcal{M}_j\mathcal{E}_j$, among the most frequent $15k$ words of both languages.

With the constructed lexica, the MPSR objective is:

$$J_r = \mathop{\mathbb{E}}_{(i,j)\sim\mathscr{L}^2} \mathop{\mathbb{E}}_{(x_i,x_j)\sim\texttt{Lex}(i,j)} L_r(\mathcal{M}_i x_i, \mathcal{M}_j x_j) \tag{3}$$

where $L_r(x, \hat{x})$ is the loss function for MPSR, for which we use the *mean square loss*. The MPSR training is depicted in Algorithm 2.

**Cross-Lingual Similarity Scaling (CSLS)** When constructing the pseudo-supervised lexica, a distance metric between embeddings is needed to compute nearest neighbors. Standard distance metrics such as the Euclidean distance or cosine similarity, however, can lead to the *hubness* problem in high-dimensional spaces when used to calculate nearest neighbors (Radovanović et al., 2010; Dinu and Baroni, 2015). Namely, some words are very likely to be the nearest neighbors of many others (hubs), while others are not the nearest neighbor of any word. This problem is addressed in the literature by designing alternative distance metrics, such as the inverted softmax (Smith et al., 2017) or the CSLS (Lample et al., 2018b). In this work, we adopt the CSLS similarity as a drop-in replacement for cosine similarity whenever a distance metric is needed. The CSLS similarity (whose negation is a distance metric) is calculated as follows:

$$\begin{aligned}\text{CSLS}(x, y) = 2\cos(x, y) \\ - \frac{1}{n} \sum_{y' \in N_Y(x)} \cos(x, y') \\ - \frac{1}{n} \sum_{x' \in N_X(y)} \cos(x', y)\end{aligned} \tag{4}$$

where $N_Y(x)$ is the set of $n$ nearest neighbors of $x$ in the vector space that $y$ comes from: $Y = \{y_1, ..., y_{|Y|}\}$, and vice versa for $N_X(y)$. In practice, we use $n = 10$.

### 3.3 Orthogonalization

As mentioned in §3, orthogonal linear mappings are the preferred choice when learning transformations between the embedding spaces of different languages (Xing et al., 2015; Smith et al., 2017). Therefore, we perform an orthogonalization update (Cisse et al., 2017) after each training step to ensure that our mappings $\mathcal{M}$ are (approximately) orthogonal:

$$\forall l : \mathcal{M}_l = (1 + \beta)\mathcal{M}_l - \beta\mathcal{M}_l\mathcal{M}_l^\top\mathcal{M}_l$$

where $\beta$ is set to $0.001$.

### 3.4 Unsupervised Multilingual Validation

In order to do model selection in the unsupervised setting, where no validation set can be used, a surrogate validation criterion is required that does not depend on bilingual data. Previous work shows promising results using such surrogate criteria for model validation in the bilingual case (Lample et al., 2018b), and we in this work adopt a variant adapted to our multilingual setting:

$$\begin{aligned}V(\mathcal{M}, \mathcal{E}) &= \mathop{\mathbb{E}}_{(i,j)\sim P_{ij}} \texttt{mean\_csls}(\mathcal{M}_j^\top\mathcal{M}_i\mathcal{E}_i, \mathcal{E}_j) \\ &= \sum_{i \neq j} p_{ij} \cdot \texttt{mean\_csls}(\mathcal{M}_j^\top\mathcal{M}_i\mathcal{E}_i, \mathcal{E}_j)\end{aligned}$$

where $p_{ij}$ forms a probability simplex. In this work, we let all $p_{ij} = \frac{1}{N(N-1)}$ so that $V(\mathcal{M}, \mathcal{E})$ reduces to the macro average over all language pairs. Using different $p_{ij}$ values can place varying weights on different language pairs, which might be desirable in certain scenarios.

The `mean_csls` function is an unsupervised bilingual validation criterion proposed by Lample et al. (2018b), which is the mean CSLS similarities between the most frequent $10k$ words and their translations (nearest neighbors).

## 4 Experiments

In this section, we present experimental results to demonstrate the effectiveness of our unsupervised MWE method on two benchmark tasks, the multilingual word translation task, and the SemEval-2017 cross-lingual word similarity task. We compare our `MAT+MPSR` method with state-of-the-art unsupervised and supervised approaches, and show that ours outperforms previous methods, supervised or not, on both tasks.

Pre-trained $300d$ fastText (monolingual) embeddings (Bojanowski et al., 2017) trained on the Wikipedia corpus are used for all systems that require monolingual word embeddings for learning cross-lingual embeddings.

### 4.1 Multilingual Word Translation

In this section, we consider the task of word translation between arbitrary pairs of a set of $N$ languages. To this end, we use the recently released multilingual word translation dataset on six languages: English, French, German, Italian, Portuguese and Spanish (Lample et al., 2018b). For any pair of the six languages, a ground-truth bilingual dictionary is provided with a train-test split of 5000 and 1500 unique source words, respectively. The $5k$ training pairs are used in training supervised baseline methods, while all unsupervised methods do not rely on any cross-lingual resources. All systems are tested on the 1500 test word pairs for each pair of languages.

For comparison, we adopted a state-of-the-art unsupervised BWE method (Lample et al., 2018b) and generalize it for the multilingual setting using the two aforementioned approaches, namely BWE-Pivot and BWE-Direct, to produce unsupervised baseline MWE systems. English is chosen as the pivot language in BWE-Pivot. We further incorporate the supervised BWE-Direct (Sup-BWE-Direct) method as a baseline, where each BWE is trained on the $5k$ gold-standard word pairs via the orthogonal Procrustes process (Artetxe et al., 2017; Lample et al., 2018b).

Table 1 presents the evaluation results, wherein the numbers represent *precision@1*, namely how many times one of the correct translations of a source word is retrieved as the top candidate. All systems retrieve word translations using the CSLS similarity in the learned embedding space. Table 1a shows the detailed results for all 30 language pairs, while Table 1b summarizes the results in a number of ways. We first observe the training cost of all systems summarized in Table 1b. #BWEs indicates the training cost of a certain method measured by how many BWE models it is equivalent to train. BWE-Pivot needs to train $2(N-1)$ BWEs since a separate BWE is trained for each direction in a language pair for increased performance. BWE-Direct on the other hand, trains an individual BWE for all (again, directed) pairs, resulting a total of $N(N-1)$ BWEs. The supervised Sup-BWE-Direct method trains the same number of BWEs as BWE-Direct but is much faster in practice, for it does not require the unsupervised adversarial training stage. Finally, while our `MAT+MPSR` method does not train independent BWEs, as argued in §3.1, the training cost is roughly equivalent to training $N-1$ BWEs, which is corroborated by the real training time shown in Table 1b.

We can see in Table 1a that our `MAT+MPSR` method achieves the highest performance on all but 3 language pairs, compared against both the unsupervised and supervised approaches. When looking at the overall performance across all language pairs, BWE-Direct achieves a $+0.6\%$ performance gain over BWE-Pivot at the cost of being much slower to train. When supervision is available, Sup-BWE-Direct further improves another $0.4\%$ over BWE-Direct. Our `MAT+MPSR` method, however, attains an impressive $1.3\%$ improvement against Sup-BWE-Direct, despite the lack of cross-lingual supervision.

To provide a more in-depth examination of the results, we first consider the Romance language pairs, such as fr-es, fr-it, fr-pt, es-it, it-pt and their reverse directions. BWE-Pivot performs notably worse than BWE-Direct on these pairs, which validates our hypothesis that going through a less similar language (English) when translating between

| | en-de | en-fr | en-es | en-it | en-pt | de-fr | de-es | de-it | de-pt | fr-es | fr-it | fr-pt | es-it | es-pt | it-pt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Supervised methods with cross-lingual supervision* | | | | | | | | | | | | | | | |
| Sup-BWE-Direct | 73.5 | 81.1 | 81.4 | 77.3 | 79.9 | 73.3 | 67.7 | 69.5 | 59.1 | 82.6 | 83.2 | 78.1 | 83.5 | 87.3 | 81.0 |
| *Unsupervised methods without cross-lingual supervision* | | | | | | | | | | | | | | | |
| BWE-Pivot | 74.0 | 82.3 | 81.7 | 77.0 | 80.7 | 71.9 | 66.1 | 68.0 | 57.4 | 81.1 | 79.7 | 74.7 | 81.9 | 85.0 | 78.9 |
| BWE-Direct | 74.0 | 82.3 | 81.7 | 77.0 | 80.7 | 73.0 | 65.7 | 66.5 | 58.5 | 83.1 | 83.0 | 77.9 | 83.3 | 87.3 | 80.5 |
| MAT+MPSR | **74.8** | **82.4** | **82.5** | **78.8** | **81.5** | **76.7** | **69.6** | **72.0** | **63.2** | **83.9** | **83.5** | **79.3** | **84.5** | **87.8** | **82.3** |
| | de-en | fr-en | es-en | it-en | pt-en | fr-de | es-de | it-de | pt-de | es-fr | it-fr | pt-fr | it-es | pt-es | pt-it |
| *Supervised methods with cross-lingual supervision* | | | | | | | | | | | | | | | |
| Sup-BWE-Direct | 72.4 | **82.4** | 82.9 | 76.9 | **80.3** | 69.5 | 68.3 | 67.5 | 63.7 | 85.8 | 87.1 | 84.3 | 87.3 | 91.5 | 81.1 |
| *Unsupervised methods without cross-lingual supervision* | | | | | | | | | | | | | | | |
| BWE-Pivot | 72.2 | 82.1 | 83.3 | **77.7** | 80.1 | 68.1 | 67.9 | 66.1 | 63.1 | 84.7 | 86.5 | 82.6 | 85.8 | 91.3 | 79.2 |
| BWE-Direct | 72.2 | 82.1 | 83.3 | **77.7** | 80.1 | 69.7 | 68.8 | 62.5 | 60.5 | 86 | 87.6 | 83.9 | 87.7 | 92.1 | 80.6 |
| MAT+MPSR | **72.9** | 81.8 | **83.7** | 77.4 | 79.9 | **71.2** | **69.0** | **69.5** | **65.7** | **86.9** | **88.1** | **86.3** | **88.2** | **92.7** | **82.6** |

(a) Detailed Results

| | *Training Cost* | | *Single Source* | | | | | | *Single Target* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | #BWEs | time | en-xx | de-xx | fr-xx | es-xx | it-xx | pt-xx | xx-en | xx-de | xx-fr | xx-es | xx-it | xx-pt | **Overall** |
| *Supervised methods with cross-lingual supervision* | | | | | | | | | | | | | | | |
| Sup-BWE-Direct | $N(N-1)$ | $4h$ | 78.6 | 68.4 | 79.2 | 81.6 | 80.0 | 80.2 | 79.0 | 68.5 | 82.3 | 82.1 | 78.9 | 77.1 | 78.0 |
| *Unsupervised methods without cross-lingual supervision* | | | | | | | | | | | | | | | |
| BWE-Pivot | $2(N-1)$ | $8h$ | 79.1 | 67.1 | 77.1 | 80.6 | 79.0 | 79.3 | **79.1** | 67.8 | 81.6 | 81.2 | 77.2 | 75.3 | 77.0 |
| BWE-Direct | $N(N-1)$ | $23h$ | 79.1 | 67.2 | 79.2 | 81.7 | 79.2 | 79.4 | **79.1** | 67.1 | 82.6 | 82.1 | 78.1 | 77.0 | 77.6 |
| MAT+MPSR | $N-1$ | $5h$ | **80.0** | **70.9** | **79.9** | **82.4** | **81.1** | **81.4** | **79.1** | **70.0** | **84.1** | **83.4** | **80.3** | **78.8** | **79.3** |

(b) Summarized Results

Table 1: Multilingual Word Translation Results for English, German, French, Spanish, Italian and Portuguese. The reported numbers are *precision@1* in percentage. All systems use the nearest neighbor under the CSLS distance for predicting the translation of a certain word.

similar languages will result in reduced accuracy. Our MAT+MPSR method, however, overcomes this disadvantage of BWE-Pivot and achieves the best performance on all these pairs through an explicit multilingual learning mechanism without increasing the computational cost.

Furthermore, our method also beats the BWE-Direct approach, which supports our second hypothesis that utilizing knowledge from languages beyond the pair itself could improve performance. For instance, there are a few pairs where BWE-Pivot outperforms BWE-Direct, such as de-it, it-de and pt-de, even though it goes through a third language (English) in BWE-Pivot. This might suggest that for some less similar language pairs, leveraging a third language as a bridge could in some cases work better than only relying on the language pair itself. German is involved in all these language pairs where BWE-Pivot outperforms than BWE-Direct, which is potentially due to the similarity between German and the pivot language English. We speculate that if choosing a different pivot language, there might be other pairs that could benefit. This observation serves as a possible explanation of the superior performance of our multilingual method over BWE-Direct, since our method utilizes knowledge from all languages during training.

### 4.2 Cross-Lingual Word Similarity

In this section, we evaluate the quality of our MWEs on the cross-lingual word similarity (CLWS) task, which assesses how well the similarity in the cross-lingual embedding space corresponds to a human-annotated semantic similarity score. The high-quality CLWS dataset from SemEval-2017 (Camacho-Collados et al., 2017) is

| | en-de | en-es | de-es | en-it | de-it | es-it | en-fa | de-fa | es-fa | it-fa | **Average** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Supervised methods with cross-lingual supervision* | | | | | | | | | | | |
| Luminoso | **.769** | **.772** | **.735** | **.787** | **.747** | **.767** | .595 | .587 | .634 | .606 | .700 |
| NASARI | .594 | .630 | .548 | .647 | .557 | .592 | .492 | .452 | .466 | .475 | .545 |
| *Unsupervised methods without cross-lingual supervision* | | | | | | | | | | | |
| BWE-Pivot | .709 | .711 | .703 | .709 | .682 | .721 | .672 | .655 | .701 | .688 | .695 |
| BWE-Direct | .709 | .711 | .703 | .709 | .675 | .726 | .672 | .662 | .714 | .695 | .698 |
| MAT+MPSR | .711 | .712 | .708 | .709 | .684 | .730 | **.680** | **.674** | **.720** | **.709** | **.704** |

Table 2: Results for the SemEval-2017 Cross-Lingual Word Similarity task. Spearman's $\rho$ is reported. Luminoso (Speer and Lowry-Duda, 2017) and NASARI (Camacho-Collados et al., 2016) are the two top-performing systems for SemEval-2017 that reported results on all language pairs.

used for evaluation. The dataset contains word pairs from any two of the five languages: English, German, Spanish, Italian, and Farsi (Persian), annotated with semantic similarity scores.

In addition to the BWE-Pivot and BWE-Direct baseline methods, we also include the two best-performing systems on SemEval-2017, Luminoso (Speer and Lowry-Duda, 2017) and NASARI (Camacho-Collados et al., 2016) for comparison. Note that these two methods are supervised, and have access to the Europarl[3] (for all languages but Farsi) and the OpenSubtitles2016[4] parallel corpora.

Table 2 shows the results, where the performance of each model is measured by the Spearman correlation. When compared to the BWE-Pivot and the BWE-Direct baselines, MAT+MPSR continues to perform the best on all language pairs. The qualitative findings stay the same as in the word translation task, except the margin is less significant. This might be because the CLWS task is much more lenient compared to the word translation task, where in the latter one needs to correctly identify the translation of a word out of hundreds of thousands of words in the vocabulary. In CLWS though, one can still achieve relatively high correlation in spite of minor inaccuracies.

On the other hand, an encouraging result is that when compared to the state-of-the-art supervised results, our MAT+MPSR method outperforms NASARI by a very large margin, and achieves top-notch overall performance similar to the competition winner, Luminoso, without using any bitexts. A closer examination reveals that our unsupervised method lags a few points behind Luminoso on the European languages wherein the supervised methods have access to the large-scale high-quality Europarl parallel corpora. It is the low-resource language, Farsi, that makes our unsupervised method stand out. All of the unsupervised methods outperform the supervised systems from SemEval-2017 on language pairs involving Farsi, which is not covered by the Europarl bitexts. This suggests the advantage of learning unsupervised embeddings for lower-resourced languages, where the supervision might be noisy or absent. Furthermore, within the unsupervised methods, MAT+MPSR again performs the best, and attains a higher margin over the baseline approaches on the low-resource language pairs, vindicating our claim of better multilingual performance.

## 5 Conclusion

In this work, we propose a fully unsupervised model for learning multilingual word embeddings (MWEs). Although methods exist for learning high-quality unsupervised BWEs (Lample et al., 2018b), little work has been done in the unsupervised multilingual setting. Previous work relies solely on a number of unsupervised BWE models to generate MWEs (e.g. BWE-Pivot and BWE-Direct), which does not fully leverage the interdependencies among all the languages. Therefore, we propose the MAT+MPSR method that explicitly exploits the relations between all language pairs without increasing the computational cost. In our experiments on multilingual word translation and cross-lingual word similarity (SemEval-2017), we show that MAT+MPSR outperforms existing unsupervised and even supervised models, achieving new state-of-the-art performance.

For future work, we plan to investigate how our

[3] http://opus.nlpl.eu/Europarl.php

[4] http://opus.nlpl.eu/OpenSubtitles2016.php

method can be extended to work with other BWE frameworks, in order to overcome the instability issue of Lample et al. (2018b). As pointed out by recent work (Søgaard et al., 2018; Artetxe et al., 2018a), the method by Lample et al. (2018b) performs much worse on certain languages such as Finnish, etc. More reliable multilingual embeddings might be obtained on these languages if we adapt our multilingual training framework to work with the more robust methods proposed recently.

## References


Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. 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, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

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.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver, Canada. Association for Computational Linguistics.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv e-prints 1606.01614v5*.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. 2017. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 854–863, International Convention Centre, Sydney, Australia. PMLR.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113, Lisbon, Portugal. Association for Computational Linguistics.

Georgiana Dinu and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *International Conference on Learning Representations, Workshop Track*.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 894–904, Valencia, Spain. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

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*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jgou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119, USA. Curran Associates Inc.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, 11:2487–2531.

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 ICLR*.

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.

Robert Speer and Joanna Lowry-Duda. 2017. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 85–89.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. 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)*, pages 719–725, Beijing, China. Association for Computational Linguistics.

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

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

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*, pages 1393–1398, Seattle, Washington, USA. Association for Computational Linguistics.