Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?

Ivan Vulić1, Goran Glavaš2, Roi Reichart3, Anna Korhonen1
1 Language Technology Lab, University of Cambridge
2 Data and Web Science Group, University of Mannheim
3 Faculty of Industrial Engineering and Management, Technion, IIT

Abstract

Recent efforts in cross-lingual word embedding (CLWE) learning have predominantly focused on fully unsupervised approaches that project monolingual embeddings into a shared cross-lingual space without any cross-lingual signal. The lack of any supervision makes such approaches conceptually attractive. Yet, their only core difference from (weakly) supervised projection-based CLWE methods is in the way they obtain a seed dictionary used to initialize an iterative self-learning procedure. The fully unsupervised methods have arguably become more robust, and their primary use case is CLWE induction for pairs of resource-poor and distant languages. In this paper, we question the ability of even the most robust unsupervised CLWE approaches to induce meaningful CLWEs in these more challenging settings. A series of bilingual lexicon induction (BLI) experiments with 15 diverse languages (210 language pairs) show that fully unsupervised CLWE methods still fail for a large number of language pairs (e.g., they yield zero BLI performance for 87/210 pairs). Even when they succeed, they never surpass the performance of weakly supervised methods (seeded with 500-1,000 translation pairs) using the same self-learning procedure in any BLI setup, and the gaps are often substantial. These findings call for revisiting the main motivations behind fully unsupervised CLWE methods.

1 Introduction and Motivation

The wide use and success of monolingual word embeddings in NLP tasks (Turian et al., 2010; Chen and Manning, 2014) has inspired further research focus on the induction of cross-lingual word embeddings (CLWEs). CLWE methods learn a shared cross-lingual word vector space where words with similar meanings obtain similar vectors regardless of their actual language. CLWEs benefit cross-lingual NLP, enabling multilingual modeling of meaning and supporting cross-lingual transfer for downstream tasks and resource-lean languages. CLWEs provide invaluable cross-lingual knowledge for, inter alia, bilingual lexicon induction (Gouws et al., 2015; Heyman et al., 2017), information retrieval (Vulić and Moens, 2015; Litschko et al., 2019), machine translation (Artetxe et al., 2018; Lample et al., 2018b), document classification (Klementiev et al., 2012), cross-lingual plagiarism detection (Glavaš et al., 2018), domain adaptation (Ziser and Reichart, 2018), cross-lingual POS tagging (Gouws and Søgaard, 2015; Zhang et al., 2016), and cross-lingual dependency parsing (Guo et al., 2015; Søgaard et al., 2015).

The landscape of CLWE methods has recently been dominated by the so-called projection-based methods (Mikolov et al., 2013a; Ruder et al., 2019; Glavaš et al., 2019). They align two monolingual embedding spaces by learning a projection/mapping based on a training dictionary of translation pairs. Besides their simple conceptual design and competitive performance, their popularity originates from the fact that they rely on rather weak cross-lingual supervision. Originally, the seed dictionaries typically spanned several thousand word pairs (Mikolov et al., 2013a; Faruqui and Dyer, 2014; Xing et al., 2015), but more recent work has shown that CLWEs can be induced with even weaker supervision from small dictionaries spanning several hundred pairs (Vulić and Korhonen, 2016), identical strings (Smith et al., 2017), or even only shared numerals (Artetxe et al., 2017).

Taking the idea of reducing cross-lingual supervision to the extreme, the latest CLWE developments almost exclusively focus on fully unsupervised approaches (Conneau et al., 2018; Artetxe et al., 2018b; Dou et al., 2018; Chen and Cardie, 2018; Alvarez-Melis and Jaakkola, 2018; Kim et al., 2018; Alaux et al., 2019; Mohiuddin and Joty, 2019, inter alia): they fully abandon any source of
(even even weak) supervision and extract the initial seed dictionary by exploiting topological similarities between pre-trained monolingual embedding spaces. Their modus operandi can roughly be described by three main components: C1) unsupervised extraction of a seed dictionary; C2) a self-learning procedure that iteratively refines the dictionary to learn projections of increasingly higher quality; and C3) a set of preprocessing and postprocessing steps (e.g., unit length normalization, mean centering, (de)whitening) (Artetxe et al., 2018a) that make the entire learning process more robust.

The induction of fully unsupervised CLWEs is an inherently interesting research topic per se. Nonetheless, the main practical motivation for developing such approaches in the first place is to facilitate the construction of multilingual NLP tools and widen the access to language technology for resource-poor languages and language pairs. However, the first attempts at fully unsupervised CLWE induction failed exactly for these use cases, as shown by Søgaard et al. (2018). Therefore, the follow-up work aimed to improve the robustness of unsupervised CLWE induction by introducing more robust self-learning procedures (Artetxe et al., 2018b; Kementchedjhieva et al., 2018). Besides increased robustness, recent work claims that fully unsupervised projection-based CLWEs can even match or surpass their supervised counterparts (Conneau et al., 2018a; Artetxe et al., 2018b; Alvarez-Melis and Jaakkola, 2018; Hoshen and Wolf, 2018; Heyman et al., 2019).

In this paper, we critically examine these claims on robustness and improved performance of unsupervised CLWEs by running a large-scale evaluation in the bilingual lexicon induction (BLI) task on 15 languages (i.e., 210 languages pairs, see Table 2 in §3). The languages were selected to represent different language families and morphological types, as we argue that fully unsupervised CLWEs have been designed to support exactly these setups. However, we show that even the most robust unsupervised CLWE method (Artetxe et al., 2018b) still fails for a large number of language pairs: 87/210 BLI setups are unsuccessful, yielding (near-)zero BLI performance. Further, even when the unsupervised method succeeds, it is because the components C2 (self-learning) and C3 (pre-/post-processing) can mitigate the undesired effects of noisy seed lexicon extraction. We then demonstrate that the combination of C2 and C3 with a small provided seed dictionary (e.g., 500 or 1K pairs) outscores the unsupervised method in all cases, often with a huge margin, and does not fail for any language pair. Furthermore, we show that the most robust unsupervised CLWE approach still fails completely when it relies on monolingual word vectors trained on domain-dissimilar corpora. We also empirically verify that unsupervised approaches cannot outperform weakly supervised approaches also for closely related languages (e.g., Swedish–Danish, Spanish–Catalan).

While the “no supervision at all” premise behind fully unsupervised CLWE methods is indeed seductive, our study strongly suggests that future research efforts should revisit the main motivation behind these methods and focus on designing even more robust solutions, given their current inability to support a wide spectrum of language pairs. In hope of boosting induction of CLWEs for more diverse and distant language pairs, we make all 210 training and test dictionaries used in this work publicly available at: https://github.com/ivulic/panlex-bli.

2 Methodology

We now dissect a general framework for unsupervised CLWE learning, and show that the “bag of tricks of the trade” used to increase their robustness (which often slips under the radar) can be equally applied to (weakly) supervised projection-based approaches, leading to their fair(er) comparison.

2.1 Projection-Based CLWE Approaches

In short, projection-based CLWE methods learn (to linearly) align independently trained monolingual spaces $X$ and $Z$, using a word translation dictionary $D_0$ to guide the alignment process. Let $X_D \subset X$ and $Z_D \subset Z$ be the row-aligned subsets of monolingual spaces containing vectors of aligned words from $D_0$. Alignment matrices $X_D$ and $Z_D$ are then used to learn orthogonal transformations $W_Z$ and $W_Z$ that define the joint bilingual space $Y = kW_Z \cup kZW_Z$. While supervised projection-based CLWE models learn the mapping using a provided external (clean) dictionary $D_0$, their unsupervised counterparts automatically induce the seed dictionary in an unsupervised way (C1) and then refine it in an iterative fashion (C2).

Unsupervised CLWEs. These methods first induce a seed dictionary $D^{(1)}$ leveraging only two unaligned monolingual spaces (C1). While the
algorithms for unsupervised seed dictionary induction differ, they all strongly rely on the assumption of similar topological structure between the two pretrained monolingual spaces. Once the seed dictionary is obtained, the two-step iterative self-learning procedure (C2) takes place: 1) a dictionary \( D^{(1)} \) is first used to learn the joint space \( Y^{(k)} = XW_x^{(k)} \cup ZW_z^{(k)} \); 2) the nearest neighbours in \( Y^{(k)} \) then form the new dictionary \( D^{(k+1)} \). We illustrate the general structure in Figure 1.

A recent empirical survey paper (Glavaš et al., 2019) has compared a variety of latest unsupervised CLWE methods (Conneau et al., 2018a; Alvarez-Melis and Jaakkola, 2018; Hoshen and Wolf, 2018; Artetxe et al., 2018b) in several downstream tasks (e.g., BLI, cross-lingual information retrieval, document classification). The results of their study indicate that the VECMAP model of Artetxe et al. (2018b) is by far the most robust and best performing unsupervised CLWE model. For the actual results and analyses, we refer the interested reader to the original paper of Glavaš et al. (2019). Another recent evaluation paper (Doval et al., 2019) as well as our own preliminary BLI tests (not shown for brevity) have further verified their findings. We thus focus on VECMAP in our analyses, and base the following description of the components C1-C3 on that model.

2.2 Three Key Components

C1. Seed Lexicon Extraction. VECMAP induces the initial seed dictionary using the following heuristic: monolingual similarity distributions for words with similar meaning will be similar across languages\(^1\). The monolingual similarity distributions for the two languages are given as rows (or columns; the matrices are symmetric) of \( M_x = XX^T \) and \( M_z = ZZ^T \). For the distributions of similarity scores to be comparable, the values in each row of \( M_x \) and \( M_z \) are first sorted. The initial dictionary \( D^{(1)} \) is finally obtained by searching for mutual nearest neighbours between the rows of \( \sqrt{M_x} \) and of \( \sqrt{M_z} \).

C2. Self-Learning. Not counting the preprocessing and postprocessing steps (component C3), self-learning then iteratively repeats two steps:

1) Let \( D^{(k)} \) be the binary matrix indicating the aligned words in the dictionary \( D^{(k)} \). The orthogonal transformation matrices are then obtained as \( W_x^{(k)} = U \) and \( W_z^{(k)} = V \), where \( U \Sigma V^T \) is the singular value decomposition of the matrix \( X^T D^{(k)} Z \). The cross-lingual space of the \( k \)-th iteration is then \( Y^{(k)} = XW_x^{(k)} \cup ZW_z^{(k)} \).

2) The new dictionary \( D^{(k+1)} \) is then built by identifying nearest neighbours in \( Y^{(k)} \). These can be easily extracted from the matrix \( P = XW_x^{(k)}(ZW_z^{(k)})^T \). All nearest neighbours can be used, or additional symmetry constraints can be imposed to extract only mutual nearest neighbours: all pairs of indices \((i, j)\) for which \( P_{ij} \) is the largest value both in row \( i \) and column \( j \).

The above procedure, however, often converges to poor local optima. To remedy for this, the second step (i.e., dictionary induction) is extended with techniques that make self-learning more robust. First, the vocabularies of \( X \) and \( Z \) are cut to the top \( k \) most frequent words\(^2\). Second, similarity scores in \( P \) are kept with probability \( p \), and set to zero otherwise. This dropout allows for a wider exploration of possible word pairs in the dictionary and contributes to escaping poor local optima given the noisy seed lexicon in the first iterations.

C3. Preprocessing and Postprocessing Steps. While iteratively learning orthogonal transformations \( W_x \) and \( W_z \) for \( X \) and \( Z \) is the central step of unsupervised projection-based CLWE methods, preprocessing and postprocessing techniques are additionally applied before and after the transformation. While such techniques are often over-

\(^1\) For instance, \( \text{zwei} \) and \( \text{two} \) will have similar distributions of similarities over their respective language vocabularies.

\(^2\) i.e., \( D^{(k)}_{ij} = 1 \iff \text{the \( i \)-th word of one language and the \( j \)-th word of the other are a translation pair in } D^{(k)} \).
looked in model comparisons, they may have a great impact on the model’s final performance, as we validate in §4. We briefly summarize two pre-processing (S1 and S2) and post-processing (S3 and S4) steps used in our evaluation, originating from the framework of Artetxe et al. (2018a).

S1) **Normalization and mean centering.** We first apply unit length normalization: all vectors in $X$ and $Z$ are normalized to have a unit Euclidean norm. Following that, $X$ and $Z$ are mean centered dimension-wise and then again length-normalized.

S2) **Whitening.** ZCA whitening (Bell and Sejnowski, 1997) is applied on (S1-processed) $X$ and $Z$: it transforms the matrices such that each dimension has unit variance and that the dimensions are uncorrelated. Intuitively, the vector spaces are easier to align along directions of high variance.

S3) **Dewhiteniing.** A transformation inverse to S2: for improved performance it is important to restore the variance information after the projection, if whitening was applied in S2 (Artetxe et al., 2018a).

S4) **Symmetric re-weighting.** This step attempts to further align the embeddings in the cross-lingual embedding space by measuring how well a dimension in the space correlates across languages for the current iteration dictionary $D^{(k)}$. The best results are obtained when re-weighting is neutral to the projection direction, that is, when it is applied symmetrically in both languages.

In the actual implementation S1 is applied only once, before self-learning. S2, S3 and S4 are applied in each self-learning iteration.

**Model Configurations.** Note that C2 and C3 can be equally used on top of any (provided) seed lexicon (i.e., $D^{(1)}:=D_0$) to enable weakly supervised learning, as we propose here. In fact, the variations of the three key components, C1 seed lexicon, C2 self-learning, and C3) preprocessing and post-processing, construct various model configurations which can be analyzed to probe the importance of each component in the CLWE induction process. A selection of representative configurations evaluated later in §4 is summarized in Table 1.

3 Experimental Setup

**Evaluation Task.** Our task is bilingual lexicon induction (BLI). It has become the de facto standard evaluation for projection-based CLWEs (Ruder et al., 2019; Glavaš et al., 2019). In short, after a shared CLWE space has been induced, the task is to retrieve target language translations for a test set of source language words. Its lightweight nature allows us to conduct a comprehensive evaluation across a large number of language pairs. Since BLI is cast as a ranking task, following Glavaš et al. (2019) we use mean average precision (MAP) as the main evaluation metric: in our BLI setup with only one correct translation for each “query” word, MAP is equal to mean reciprocal rank (MRR).

(Selection of) **Language Pairs.** Our selection of test languages is guided by the following goals: a) following recent initiatives in other NLP research (e.g., for language modeling) (Cotterell et al., 2018; Gerz et al., 2018), we aim to ensure the coverage of different genealogical and typological language properties, and b) we aim to analyze a large set of language pairs and offer new evaluation data which extends and surpasses other work in the CLWE literature. These two properties will facilitate analyses between (dis)similar language pairs and offer a comprehensive set of evaluation setups that test the robustness and portability of fully unsupervised CLWEs. The final list of 15 diverse test languages is provided in Table 2, and includes samples from different languages types and families. We run BLI evaluations for all language pairs in both directions, for a total of $15 \times 14=210$ BLI setups.

**Monolingual Embeddings.** We use the 300-dim vectors of Grave et al. (2018) for all 15 languages, pretrained on Common Crawl and Wikipedia with fastText (Bojanowski et al., 2017). We trim all

---

3More formally, assume that we are working with matrices $X$ and $Z$ that already underwent all transformations described in S1-S3. Another matrix $D$ represents the current bilingual dictionary $D$. $D_{ij} = 1$ if the $i^{th}$ source word is translated by the $j^{th}$ target word and $D_{ij} = 0$ otherwise. Then, given the singular value decomposition $USV^T = X^TDZ$, the final re-weighted projection matrices are $W_s = US\frac{1}{2}$ and $W_z = VS\frac{1}{2}$. We refer the reader to (Artetxe et al., 2018a) and (Artetxe et al., 2018b) for more details.

5We aim to analyze whether a CLWE method manages to learn anything meaningful at all, and can indicate “unsuccessful” CLWE induction (e.g., when BLI performance is similar to a random baseline): detecting such CLWEs is especially important when dealing with fully unsupervised models.

7We trim all
We now summarize and discuss the main findings provided in Table 1 that show the same trends in the final results. As in prior work (Conneau et al., 2018a; Glavaš et al., 2019), we then reserve the 5K pairs created from the more frequent 5K words for training, while the next 2K pairs are used for test. Smaller training dictionaries (1K and 500 pairs) are created by again selecting pairs comprising the most frequent 5K words.

Training Setup. In all experiments, we set the hyper-parameters to values that were tuned in prior research. When extracting the unsupervised seed lexicon, the 4K most frequent words of each language are used; self-learning operates on the 20K most frequent words of each language; with dropout the keep probability p is 0.1: CSLS with k = 10 nearest neighbors (Artetxe et al., 2018b).

Again, Table 1 lists the main model configurations in our comparison. For the fully unsupervised model we always report the best performing configuration after probing different self-learning strategies (i.e., +SL, +SL+NOD, and +SL+SYM are tested). The results for unsupervised are always reported as averages over 5 restarts: this means that with unsupervised we count BLI setups as unsuccessful only if the results are close to zero in all 5/5 runs. ORTHG-SUPER is the standardized supervised model with orthogonal projections from prior work (Smith et al., 2017; Glavaš et al., 2019).

4 Results and Discussion

Main BLI results averaged over each source language (L₁) are provided in Table 3 and Table 4. We now summarize and discuss the main findings across several dimensions of comparison.

Unsupervised vs. (Weakly) Supervised. First, when exactly the same components C2 and C3 are used, unsupervised is unable to outperform a (weakly) supervised full+SL+SYM variant, and the gap in final performance is often substantial. In fact, full+SL+SYM and full+SL+NOD outperform the best unsupervised for all 210/210 BLI setups: we observe the same phenomenon with varying dictionary sizes, that is, it equally holds when we seed self-learning with 5K, 1K, and 500 translation pairs, see also Figure 2. This also suggests that the main reason why unsupervised approaches were considered on-par with supervised approaches in prior work (Conneau et al., 2018a; Artetxe et al., 2018b) is because they were not compared under fair circumstances: while unsupervised...
Table 3: BLI scores (MRR) for all model configurations. The scores are averaged over all experimental setups where each of the 15 languages is used as \(L_1\); e.g., CA-* means that the translation direction is from Catalan (CA) as source (\(L_1\)) to each of the remaining 14 languages listed in Table 2 as targets (\(L_2\)), and we average over the corresponding 14 CA-* BLI setups. \(5k\) and \(1k\) denote the seed dictionary size for (weakly) supervised methods \((D_0)\). Unsuccessful setups refer to the number of BLI experimental setups with the fully UNSUPERVISED model that yield an MRR score \(\leq 0.01\). The \textbf{Avg} column refers to the averaged MRR scores of each model configuration over all 15×14=210 BLI setups. The highest scores for two different seed dictionary sizes in each column are in bold, the second best are underlined. See Table 1 for the brief description of all model configurations in the comparison. Full results for each particular language pair are available in the supplemental material.

Table 4: Summary statistics computed over all 15×14=210 BLI setups. \textbf{a)} \textbf{Unsucc.} denotes the total number of unsuccessful setups, where a setup is considered unsuccessful if MRR \(\leq 0.01\) or MRR \(\leq 0.05\) (in the parentheses); \textbf{b)} \textbf{Win} refers to the total number of “winning” setups, that is, for all language pairs it counts how many times each particular model yields the best overall MRR score. We compute separate statistics for two settings \(|D_0| = 1k\) and \(|D_0| = 5k\).

Table 3: BLI scores (MRR) for all model configurations. The scores are averaged over all experimental setups where each of the 15 languages is used as \(L_1\); e.g., CA-* means that the translation direction is from Catalan (CA) as source (\(L_1\)) to each of the remaining 14 languages listed in Table 2 as targets (\(L_2\)), and we average over the corresponding 14 CA-* BLI setups. \(5k\) and \(1k\) denote the seed dictionary size for (weakly) supervised methods \((D_0)\). Unsuccessful setups refer to the number of BLI experimental setups with the fully UNSUPERVISED model that yield an MRR score \(\leq 0.01\). The \textbf{Avg} column refers to the averaged MRR scores of each model configuration over all 15×14=210 BLI setups. The highest scores for two different seed dictionary sizes in each column are in bold, the second best are underlined. See Table 1 for the brief description of all model configurations in the comparison. Full results for each particular language pair are available in the supplemental material.

Table 4: Summary statistics computed over all 15×14=210 BLI setups. \textbf{a)} \textbf{Unsucc.} denotes the total number of unsuccessful setups, where a setup is considered unsuccessful if MRR \(\leq 0.01\) or MRR \(\leq 0.05\) (in the parentheses); \textbf{b)} \textbf{Win} refers to the total number of “winning” setups, that is, for all language pairs it counts how many times each particular model yields the best overall MRR score. We compute separate statistics for two settings \(|D_0| = 1k\) and \(|D_0| = 5k\).

Table 3: BLI scores (MRR) for all model configurations. The scores are averaged over all experimental setups where each of the 15 languages is used as \(L_1\); e.g., CA-* means that the translation direction is from Catalan (CA) as source (\(L_1\)) to each of the remaining 14 languages listed in Table 2 as targets (\(L_2\)), and we average over the corresponding 14 CA-* BLI setups. \(5k\) and \(1k\) denote the seed dictionary size for (weakly) supervised methods \((D_0)\). Unsuccessful setups refer to the number of BLI experimental setups with the fully UNSUPERVISED model that yield an MRR score \(\leq 0.01\). The \textbf{Avg} column refers to the averaged MRR scores of each model configuration over all 15×14=210 BLI setups. The highest scores for two different seed dictionary sizes in each column are in bold, the second best are underlined. See Table 1 for the brief description of all model configurations in the comparison. Full results for each particular language pair are available in the supplemental material.

Table 4: Summary statistics computed over all 15×14=210 BLI setups. \textbf{a)} \textbf{Unsucc.} denotes the total number of unsuccessful setups, where a setup is considered unsuccessful if MRR \(\leq 0.01\) or MRR \(\leq 0.05\) (in the parentheses); \textbf{b)} \textbf{Win} refers to the total number of “winning” setups, that is, for all language pairs it counts how many times each particular model yields the best overall MRR score. We compute separate statistics for two settings \(|D_0| = 1k\) and \(|D_0| = 5k\).

| \(|D_0| = 5k\) | \(|D_0| = 1k\) |
|---|---|
| Unsucc. | Win | Unsucc. | Win |
| UNSUPERVISED | 87 (94) | 0 | 87 (94) | 0 |
| ORTHG-SUPER | 0 (2) | 0 | 2 (82) | 0 |
| ORTHG+SL+SYM | 0 (1) | 0 | 1 (34) | 0 |
| FULL-SUPER | 0 (0) | 46 | 0 (41) | 0 |
| FULL+SL | 0 (7) | 0 | 0 (9) | 0 |
| FULL+SL+NON | 0 (1) | 7 | 0 (3) | 33 |
| FULL+SL+SYM | 0 (0) | 157 | 0 (0) | 177 |

| \(|D_0| = 5k\) | \(|D_0| = 1k\) |
|---|---|
| Unsucc. | Win | Unsucc. | Win |
| UNSUPERVISED | 87 (94) | 0 | 87 (94) | 0 |
| ORTHG-SUPER | 0 (2) | 0 | 2 (82) | 0 |
| ORTHG+SL+SYM | 0 (1) | 0 | 1 (34) | 0 |
| FULL-SUPER | 0 (0) | 46 | 0 (41) | 0 |
| FULL+SL | 0 (7) | 0 | 0 (9) | 0 |
| FULL+SL+NON | 0 (1) | 7 | 0 (3) | 33 |
| FULL+SL+SYM | 0 (0) | 157 | 0 (0) | 177 |

VISEd relied heavily on the components C2 and C3, these were omitted when running supervised baselines. Our unbiased comparison reveals that there is a huge gap even when supervised projection-based approaches consume only several hundred translation pairs to initiate self-learning. Are UNSUPERVISED CLWEs Robust? The results also indicate that, contrary to the beliefs established by very recent work (Artetxe et al., 2018b; Mohiuddin and Joty, 2019), fully UNSUPERVISED approaches are still prone to getting stuck in local optima, and still suffer from robustness issues when dealing with distant language pairs: 87 out of 210 BLI setups (41.4%) result in (near-)zero BLI performance, see also Table 4. At the same time, weakly supervised methods with a seed lexicon of 1k or 500 pairs do not suffer from the robustness problem and always converge to a good solution,
Figure 2: A comparison of average BLI scores with different seed dictionary sizes $D_0$ between a fully unsupervised method (UNSUPER), a supervised method without self-learning (SUPER), and two best performing weakly supervised methods with self learning (+SL+NOD and +SL+SYM). While SUPER without self-learning displays a steep drop in performance with smaller seed dictionaries, there is only a slight decrease when self-learning is used: e.g., 500 translation pairs are still sufficient to initialize robust self-learning.

as also illustrated by the results reported in Table 5.

**How Important are Preprocessing and Post-processing?** The comparisons between ORTHG-SUPER (and ORTHG+SL+SYM) on the one hand, and FULL-SUPER (and FULL+SL+SYM) on the other hand clearly indicate that the component C3 plays a substantial role in effective CLWE learning. FULL-SUPER, which employs all steps S1-S4 (see §2), outperforms ORTHG-SUPER in 208/210 setups with $|D_0|=5k$ and in 210/210 setups with $|D_0|=1k$. Similarly, FULL+SL+SYM is better than ORTHG+SL+SYM in 210/210 setups (both for $|D_0|=1k, 5k$). The scores also indicate that dropout with self-learning is useful only when we work with noisy unsupervised seed lexicons: FULL+SL+NOD and FULL+SL+SYM without dropout consistently outperform FULL+SL across the board.

**How Important is (Robust) Self-Learning?** We note that the best self-learning method is often useful even when $|D_0|=5k$ (i.e., FULL+SL+SYM is better than FULL-SUPER in 164/210 setups). However, the importance of robust self-learning gets more pronounced as we decrease the size of $D_0$: FULL+SL+SYM is better than FULL-SUPER in 210/210 setups when $|D_0|=500$ or $|D_0|=1,000$. The gap between the two models, as shown in Figure 2, increases dramatically in favor of FULL+SL+SYM as we decrease $|D_0|$.

Again, just comparing FULL-SUPER and UNSUPERVISED in Figure 2 might give a false impression that fully unsupervised CLWE methods can match their supervised counterparts, but the comparison to FULL+SL+SYM reveals the true extent of performance drop when we abandon even weak supervision. The scores also reveal that the choice of self-learning (C2) does matter: all best performing BLI runs with $|D_0|=1k$ are obtained by two configs with self-learning, and FULL+SL+SYM is the best configuration for 177/210 setups (see Table 4).

**Language Pairs.** As suggested before by Søgaard et al. (2018) and further verified by Glavaš et al. (2019) and Doval et al. (2019), the language pair at hand can have a huge impact on CLWE induction: the adversarial method of Conneau et al. (2018a) often gets stuck in poor local optima and yields degenerate solutions for distant language pairs such as English-Finnish. More recent CLWE methods (Artetxe et al., 2018b; Mohiuddin and Joty, 2019) focus on mitigating this robustness issue. However, they still rely on one critical assumption which leads them to degraded performance for distant language pairs: they assume approximate isomorphism (Barone, 2016; Søgaard et al., 2018) between monolingual embedding spaces to learn the initial seed dictionary. In other words, they assume very similar geometric constellations between two monolingual spaces: due to the Zipfian phenomena in language (Zipf, 1949) such near-isomorphism can be satisfied only for similar languages and for similar domains used for training monolingual vectors. This property is reflected in the results reported in Table 3, the number of unsuccessful setups in Table 4, as well as later in Figure 4.

For instance, the largest number of unsuccessful BLI setups with the UNSUPERVISED model is reported for Korean, Thai (a tonal language), and Basque (a language isolate): their morphological and genealogical properties are furthest away from other languages in our comparison. A substantial number of unsuccessful setups is also observed with other two language outliers from our set (see Table 2 again), Georgian and Indonesian, as well as with morphologically-rich languages such as Estonian or Turkish.

One setting in which fully UNSUPERVISED methods did show impressive results in prior work are similar language pairs. However, even in these settings when the comparison to the weakly supervised FULL-SUPER+SYM is completely fair (i.e.,
the same components C2 and C3 are used for both), UNSUPERVISED still falls short of FULL-SUPER+SYM. These results for three source languages are summarized in Figure 3. What is more, one could argue that we do not need unsupervised CLWEs for similar languages in the first place: we can harvest cheap supervision here, e.g., cognates. The main motivation behind unsupervised approaches is to support dissimilar and resource-poor language pairs for which supervision cannot be guaranteed.

**Domain Differences.** Finally, we also verify that UNSUPERVISED CLWEs still cannot account for domain differences when training monolingual vectors. We rely on the probing test of Søgaard et al. (2018): 300-dim fastText vectors are trained on 1.1M sentences on three corpora: 1) EuroParl.v7 (Koehn, 2005) (parliamentary proceedings); 2) Wikipedia (Al-Rfou et al., 2013), and 3) EMEA (Tiedemann, 2009) (medical), and BLI evaluation for three language pairs is conducted on standard MUSE BLI test sets (Conneau et al., 2018a). The results, summarized in Figure 4, reveal that UNSUPERVISED methods are able to yield a good solution only when there is no domain mismatch and for the pair with two most similar languages (English-Spanish), again questioning their robustness and portability to truly low-resource and more challenging setups. Weakly supervised methods (|D₀| = 500 or D₀ seeded with identical strings), in contrast, yield good solutions for all setups.

### 5 Further Discussion and Conclusion

The superiority of weakly supervised methods (e.g., FULL-SL-SYM) over unsupervised methods is especially pronounced for distant and typologi-
cally heterogeneous language pairs. However, our study also indicates that even carefully engineered projection-based methods with some seed supervision yield lower absolute performance for such pairs. While we have witnessed the proliferation of fully unsupervised CLWE models recently, some fundamental questions still remain. For instance, the underlying assumption of all projection-based methods (both supervised and unsupervised) is the topological similarity between monolingual spaces, which is why standard simple linear projections result in lower absolute BLI scores for distant pairs (see Table 4 and results in the supplemental material). Unsupervised approaches even exploit the assumption twice as their seed extraction is fully based on the topological similarity.

Future work should move beyond the restrictive assumption by exploring new methods that can, e.g., 1) increase the isomorphism between monolingual spaces (Zhang et al., 2019) by distinguishing between language-specific and language-pair-invariant subspaces; 2) learn effective non-linear or multiple local projections between monolingual spaces similar to the preliminary work of Nakashole (2018); 3) similar to Vulić and Korhonen (2016) and Lubin et al. (2019) “denoisify” seed lexicons during the self-learning procedure. For instance, keeping only mutual/symmetric nearest neighbour as in FULL+SL+SYM can be seen as a form of rudimentary denoising: it is indicative to see that the best overall performance in this work is reported with that model configuration.

Further, the most important contributions of unsupervised CLWE models are, in fact, the improved and more robust self-learning procedures (component C2) and technical enhancements (component C3). In this work we have demonstrated that these components can be equally applied to weakly supervised approaches: starting from a set of only several hundred pairs, they can guarantee consistently improved performance across the board. As there is still no clear-cut use case scenario for unsupervised CLWEs, instead of “going fully unsupervised”, one pragmatic approach to widen the scope of CLWE learning and its application might invest more effort into extracting at least some seed supervision for a variety of language pairs (Artetxe et al., 2017). This finding aligns well with the ongoing initiatives of the PanLex project (Kamholz et al., 2014) and the ASJP database (Wichmann et al., 2018), which aim to collate at least some translation pairs in most of the world’s languages.

Finally, this paper demonstrates that, in order to enable fair comparisons, future work on CLWEs should focus on evaluating the CLWE methods’ constituent components (e.g., components C1-C3 from this work) instead of full-blown composite systems directly. One goal of the paper is to acknowledge that the work on fully unsupervised CLWE methods has indeed advanced state-of-the-art in cross-lingual word representation learning by offering new solutions also to weakly supervised CLWE methods. However, the robustness problems are still prominent with fully unsupervised CLWEs, and future work should invest more time and effort into developing more robust and more effective methods, e.g., by reaching beyond projection-based methods towards joint approaches (Ruder et al., 2019; Ormazabal et al., 2019).

Acknowledgments

This work is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909). The work of Goran Glavaš is supported by the Baden-Württemberg Stiftung (AGREE grant of the Eliteprogramm). Roi Reichart is partially funded by ISF personal grants No. 1625/18. We thank the three anonymous reviewers for their encouraging comments and suggestions.

References


E.g., unsupervised CLWEs are fully substitutable with the superior weakly supervised CLWEs in unsupervised NMT architectures (Artetxe et al., 2018c; Lample et al., 2018a,b), or in domain adaptation systems (Ziser and Reichart, 2018) and fully unsupervised cross-lingual IR (Litschko et al., 2018).

https://asjp.clld.org/


