Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?

Ivan Vulić[1], **Goran Glavaš**[2], **Roi Reichart**[3], **Anna Korhonen**[1]
[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
{iv250,alk23}@cam.ac.uk
goran@informatik.uni-mannheim.de roiri@ie.technion.ac.il

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., 2018c; 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., 2018a; 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

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4407–4418,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

(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 $\boldsymbol{X}$ and $\boldsymbol{Z}$, using a word translation dictionary D_0 to guide the alignment process. Let $\boldsymbol{X}_D \subset \boldsymbol{X}$ and $\boldsymbol{Z}_D \subset \boldsymbol{Z}$ be the row-aligned subsets of monolingual spaces containing vectors of aligned words from D_0. Alignment matrices $\boldsymbol{X}_D$ and $\boldsymbol{Z}_D$ are then used to learn orthogonal transformations $\boldsymbol{W}_x$ and $\boldsymbol{W}_z$ that define the joint bilingual space $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}_x \cup \boldsymbol{Z}\boldsymbol{W}_z$. 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

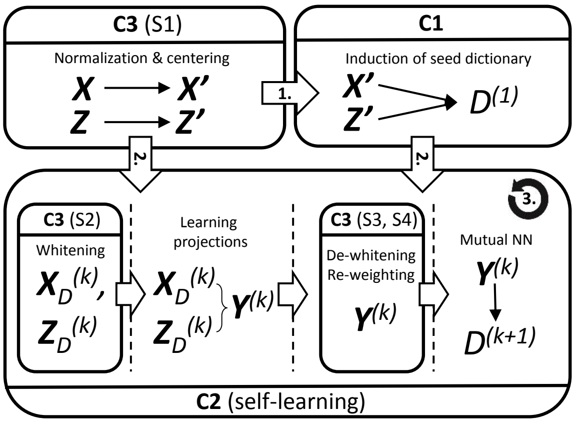


Figure 1: General unsupervised CLWE approach.

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 $D^{(k)}$ is first used to learn the joint space $\boldsymbol{Y}^{(k)} = \boldsymbol{X}\boldsymbol{W}_x^{(k)} \cup \boldsymbol{Z}\boldsymbol{W}_z^{(k)}$; 2) the nearest neighbours in $\boldsymbol{Y}^{(k)}$ 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 $\boldsymbol{M}_x = \boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{M}_z = \boldsymbol{Z}\boldsymbol{Z}^T$. For the distributions of similarity scores to be comparable, the values in each row of $\boldsymbol{M}_x$ and $\boldsymbol{M}_z$ are first sorted. The initial dictionary $D^{(1)}$ is finally obtained by searching for mutual nearest neighbours between the rows of $\sqrt{\boldsymbol{M}_x}$ and of $\sqrt{\boldsymbol{M}_z}$.

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

1) Let $\boldsymbol{D}^{(k)}$ be the binary matrix indicating the aligned words in the dictionary $D^{(k)}$.[2] The orthogonal transformation matrices are then obtained as $\boldsymbol{W}_x^{(k)} = \boldsymbol{U}$ and $\boldsymbol{W}_z^{(k)} = \boldsymbol{V}$, where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is the singular value decomposition of the matrix $\boldsymbol{X}^T\boldsymbol{D}^{(k)}\boldsymbol{Z}$. The cross-lingual space of the k-th iteration is then $\boldsymbol{Y}^{(k)} = \boldsymbol{X}\boldsymbol{W}_x^{(k)} \cup \boldsymbol{Z}\boldsymbol{W}_z^{(k)}$.

2) The new dictionary $D^{(k+1)}$ is then built by identifying nearest neighbours in $\boldsymbol{Y}^{(k)}$. These can be easily extracted from the matrix $\boldsymbol{P} = \boldsymbol{X}\boldsymbol{W}_x^{(k)}(\boldsymbol{Z}\boldsymbol{W}_z^{(k)})^T$. *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 $\boldsymbol{P}_{ij}$ 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 $\boldsymbol{X}$ and $\boldsymbol{Z}$ are cut to the top k most frequent words.[3] Second, similarity scores in $\boldsymbol{P}$ 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 $\boldsymbol{W}_x$ and $\boldsymbol{W}_z$ for $\boldsymbol{X}$ and $\boldsymbol{Z}$ 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, *zwei* and *two* will have similar distributions of similarities over their respective language vocabularies – *zwei* is expected to be roughly as (dis)similar to *drei* and *Katze* as *two* is to *three* and *cat*.

[2] I.e., $\boldsymbol{D}_{ij}^{(k)} = 1 \iff$ the i-th word of one language and the j-th word of the other are a translation pair in $D^{(k)}$.

[3] This is done to prevent spurious nearest neighbours consisting of infrequent words with unreliable vectors.

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 preprocessing (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 $\boldsymbol{X}$ and $\boldsymbol{Z}$ are normalized to have a unit Euclidean norm. Following that, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are mean centered dimension-wise and then again length-normalized.

S2) *Whitening*. ZCA whitening (Bell and Sejnowski, 1997) is applied on (S1-processed) $\boldsymbol{X}$ and $\boldsymbol{Z}$: 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) *Dewhitening*. 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 $D^{(k)}$.[4] 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 postprocessing, 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.[5] 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).[6]

(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 15×14=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).[7] We trim all

[4] More formally, assume that we are working with matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ that already underwent all transformations described in S1-S3. Another matrix $\boldsymbol{D}$ 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 $\boldsymbol{USV^T} = \boldsymbol{X^TDZ}$, the final re-weighted projection matrices are $\boldsymbol{W}_x = \boldsymbol{US}^{\frac{1}{2}}$ (and $\boldsymbol{W}_z = \boldsymbol{VS}^{\frac{1}{2}}$. We refer the reader to (Artetxe et al., 2018a) and (Artetxe et al., 2018b) for more details.

[5] While BLI is an intrinsic task, as discussed by Glavaš et al. (2019) it is a strong indicator of CLWE quality also for downstream tasks: relative performance in the BLI task correlates well with performance in cross-lingual information retrieval (Litschko et al., 2018) or natural language inference (Conneau et al., 2018b). More importantly, it also provides a means 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.

[6] MRR is more informative than the more common *Precision@1 (P@1)*; our main findings are also valid when P@1 is used; we do not report the results for brevity.

[7] Experiments with other monolingual vectors such as the

Configuration	C1	C2	C3
UNSUPERVISED	unsupervised	all tested, we always report the best one	S1-S4 (FULL)
ORTHG-SUPER	provided	–	length normalization only (partial S1)
ORTHG+SL+SYM	provided	symmetric: mutual nearest neighbours	length normalization only (partial S1)
FULL-SUPER	provided	–	S1-S4 (FULL)
FULL+SL	provided	(Artetxe et al., 2018b) with dropout	S1-S4 (FULL)
FULL+SL+NOD	provided	(Artetxe et al., 2018b) w/o dropout	S1-S4 (FULL)
FULL+SL+SYM	provided	symmetric: mutual nearest neighbours, w/o dropout	S1-S4 (FULL)

Table 1: Configurations obtained by varying components C1, C2, and C3 used in our empirical comparison in §4.

Language	Family	Type	ISO 639-1
Bulgarian	IE: Slavic	fusional	BG
Catalan	IE: Romance	fusional	CA
Esperanto	– (constructed)	agglutinative	EO
Estonian	Uralic	agglutinative	ET
Basque	– (isolate)	agglutinative	EU
Finnish	Uralic	agglutinative	FI
Hebrew	Afro-Asiatic	introflexive	HE
Hungarian	Uralic	agglutinative	HU
Indonesian	Austronesian	isolating	ID
Georgian	Kartvelian	agglutinative	KA
Korean	Koreanic	agglutinative	KO
Lithuanian	IE: Baltic	fusional	LT
Bokmål	IE: Germanic	fusional	NO
Thai	Kra-Dai	isolating	TH
Turkish	Turkic	agglutinative	TR

Table 2: The list of 15 languages from our main BLI experiments along with their corresponding language family (IE = Indo-European), broad morphological type, and their ISO 639-1 code.

vocabularies to the 200K most frequent words.

Training and Test Dictionaries. They are derived from PanLex (Baldwin et al., 2010; Kamholz et al., 2014), which was used in prior work on cross-lingual word embeddings (Duong et al., 2016; Vulić et al., 2017). PanLex currently spans around 1,300 language varieties with over 12M expressions: it offers some support and supervision also for low-resource language pairs (Adams et al., 2017). For each source language (L_1), we automatically translate their vocabulary words (if they are present in PanLex) to all 14 target (L_2) languages. To ensure the reliability of the translation pairs, we retain only unigrams found in the vocabularies of the respective L_2 monolingual spaces which scored above a PanLex-predefined threshold.

As in prior work (Conneau et al., 2018a; Glavaš et al., 2019), we then reserve the 5K pairs created from the more frequent L_1 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 L_1 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 standard 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 (L_1) 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 UNSUPER-

original fastText and skip-gram (Mikolov et al., 2013b) trained on Wikipedia show the same trends in the final results.

	BG-*	CA-*	EO-*	ET-*	EU-*	FI-*	HE-*	HU-*	ID-*	KA-*	KO-*	LT-*	NO-*	TH-*	TR-*	**Avg**
UNSUPERVISED	0.208	0.224	0.128	0.155	0.036	0.181	0.186	0.206	0.110	0.106	0.001	0.179	0.239	0.000	0.133	0.140
Unsuccessful setups	*3/14*	*2/14*	*3/14*	*6/14*	*10/14*	*4/14*	*2/14*	*3/14*	*7/14*	*6/14*	*14/14*	*4/14*	*3/14*	*14/14*	*6/14*	*87/210*
5K:ORTHG-SUPER	0.258	0.237	0.201	0.210	0.151	0.233	0.198	0.259	0.199	0.163	0.154	0.194	0.250	0.109	0.207	0.201
5K:ORTHG+SL+SYM	0.281	0.264	0.219	0.225	0.164	0.256	0.217	0.283	0.216	0.177	0.166	0.212	0.273	0.117	0.226	0.220
5K:FULL-SUPER	0.343	0.335	0.304	0.301	0.228	0.324	0.287	0.354	0.261	0.250	0.239	0.302	0.332	**0.154**	0.283	0.286
5K:FULL+SL	0.271	0.262	0.240	0.236	0.161	0.260	0.217	0.282	0.180	0.191	0.152	0.217	0.274	0.056	0.204	0.214
5K:FULL+SL+NOD	0.316	0.311	0.295	0.276	0.204	0.320	0.260	0.330	0.220	0.229	0.207	0.272	0.318	0.106	0.253	0.261
5K:FULL+SL+SYM	**0.361**	**0.356**	**0.336**	**0.316**	**0.244**	**0.348**	**0.294**	**0.374**	**0.272**	**0.263**	**0.251**	**0.310**	**0.356**	0.148	**0.299**	**0.302**
1K:ORTHG-SUPER	0.104	0.088	0.065	0.082	0.049	0.088	0.066	0.101	0.069	0.050	0.040	0.067	0.099	0.034	0.068	0.071
1K:ORTHG+SL+SYM	0.203	0.167	0.106	0.157	0.079	0.168	0.133	0.191	0.119	0.092	0.063	0.135	0.186	0.052	0.129	0.132
1K:FULL-SUPER	0.146	0.129	0.098	0.117	0.065	0.117	0.096	0.143	0.089	0.079	0.061	0.111	0.127	0.044	0.091	0.101
1K:FULL+SL	0.268	0.260	0.238	0.232	0.158	0.257	0.217	0.279	0.180	0.185	0.148	0.220	0.274	0.054	0.204	0.212
1K:FULL+SL+NOD	0.312	0.307	0.284	0.272	0.197	0.311	0.255	0.327	0.216	0.223	0.197	0.269	0.315	0.096	0.248	0.255
1K:FULL+SL+SYM	**0.341**	**0.327**	**0.302**	**0.293**	**0.212**	**0.329**	**0.268**	**0.354**	**0.243**	**0.237**	**0.203**	**0.284**	**0.337**	**0.103**	**0.274**	**0.274**

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 ≤ 0.01. The **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.

	$\lvert D_0\rvert = 5k$		$\lvert D_0\rvert = 1k$	
	Unsuc.	**Win**	**Unsuc.**	**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+NOD	0 (1)	7	0 (3)	33
FULL+SL+SYM	0 (0)	157	0 (0)	177

Table 4: Summary statistics computed over all 15×14=210 BLI setups. **a) Unsuc.** 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); **b) 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 $|D_0| = 5k$).

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,

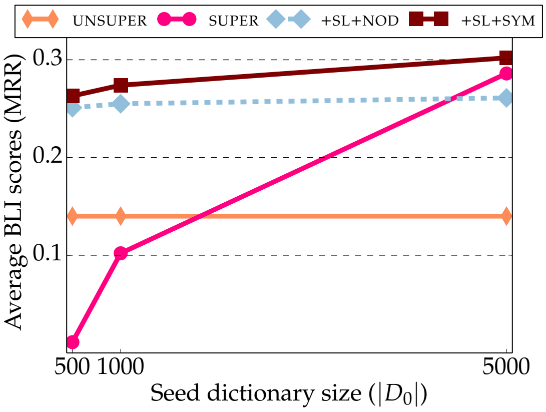


Figure 2: A comparison of average BLI scores with different seed dictionary sizes D_0 between a fully unsupervised method (UNSUPER), a supervised method without self-learning (SUPER), and two best performing weakly supervised methods with self learning (+SL+NOD and +SL+SYM). While SUPER without self-learning 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 Postprocessing? 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).

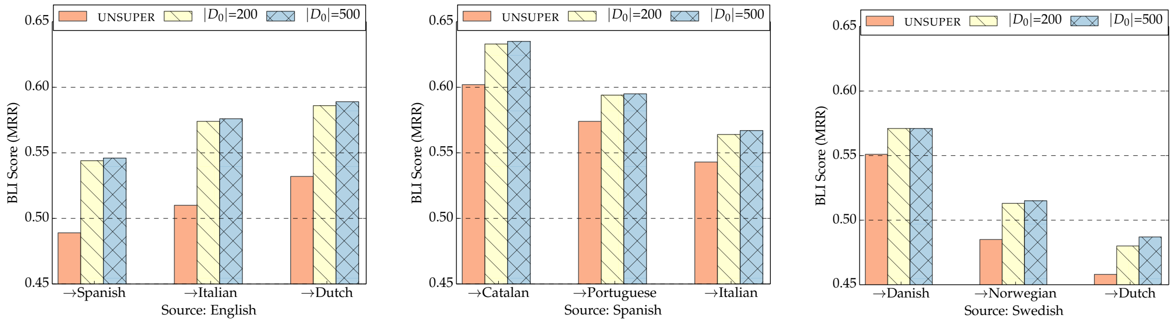
Language Pairs. As suggested before by Søgaard et al. (2018) and further verified by Glavaš et al. (2019) and Doval et al. (2019), the language pair at hand can have a huge impact on CLWE induction: the adversarial method of Conneau et al. (2018a) often gets stuck in poor local optima and yields degenerate 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.,

	BG-EU	EU-TR	FI-KO	ID-ET	ID-TH	KA-FI	KA-ID	KO-TR	NO-EU	TR-TH
UNSUPERVISED	0.005	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000
1K:FULL+SL+SYM	**0.279**	**0.212**	**0.211**	**0.213**	**0.226**	**0.306**	**0.155**	**0.279**	**0.300**	**0.137**
500:FULL+SL+SYM	0.245	0.189	0.192	0.195	0.188	0.285	0.138	0.264	0.266	0.109

Table 5: Results for a selection of BLI setups which were unsuccessful with the UNSUPERVIED CLWE method.

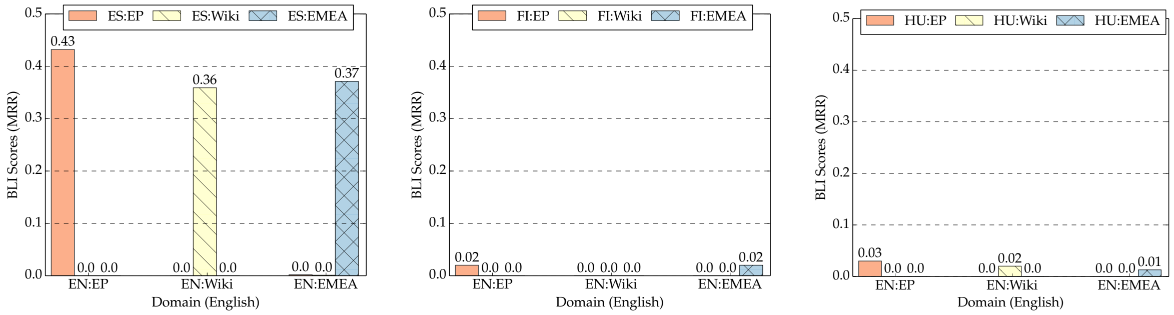


(a) English→ L_2

(b) Spanish→ L_2

(c) Swedish→ L_2

Figure 3: A comparison of BLI scores on "easy" (i.e., similar) language pairs between the fully UNSUPERVISED model and a weakly supervised model (seed dictionary size $|D_0| = 200$ or $|D_0| = 500$) which relies on the self-learning procedure with the symmetry constraint (FULL+SL+SYM).



(a) English→ Spanish

(b) English → Finnish

(c) English → Hungarian

Figure 4: BLI scores with the (most robust) fully UNSUPERVISED model for different language pairs when monolingual word embeddings are trained on dissimilar domains: parliamentary proceedings (EuroParl), Wikipedia (Wiki), and medical corpora (EMEA). Training and test data are the same as in (Søgaard et al., 2018).

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 ($|D_0| = 500$ or D_0 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 typologically 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 denoisifying: 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,[8] 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.[9]

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

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of EACL*, pages 937–947.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of CoNLL*, pages 183–192.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyperalignment for multilingual word embeddings. In *Proceedings of ICLR*.

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of EMNLP*, pages 1881–1890.

[8] 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).

[9] https://asjp.clld.org/

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of AAAI*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pages 789–798.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *Proceedings of ICLR*.

Timothy Baldwin, Jonathan Pool, and Susan Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Proceedings of COLING (Demo Papers)*, pages 37–40.

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

Anthony Bell and Terrence Sejnowski. 1997. The 'Independent Components' of Natural Scenes are Edge Filters. *Vision Research*.

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

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*, pages 740–750.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of EMNLP*, pages 261–270.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *Proceedings of ICLR*.

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 EMNLP*, pages 2475–2485.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of NAACL-HLT*.

Zi-Yi Dou, Zhi-Hao Zhou, and Shujian Huang. 2018. Unsupervised bilingual lexicon induction via latent variable models. In *Proceedings of EMNLP*, pages 621–626.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2019. On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning. *CoRR*, abs/1908.07742.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of EMNLP*, pages 1285–1295.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*, pages 462–471.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of EMNLP*, pages 316–327.

Goran Glavaš, Marc Franco-Salvador, Simone P Ponzetto, and Paolo Rosso. 2018. A resource-light method for cross-lingual semantic textual similarity. *Knowledge-Based Systems*, 143:1–9.

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 ACL*, pages 710–721.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML*, pages 748–756.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of NAACL-HLT*, pages 1386–1390.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC*, pages 3483–3487.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of ACL*, pages 1234–1244.

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. In *Proceedings of NAACL-HLT*, pages 1890–1902.

Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2017. Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proceedings of EACL*, pages 1085–1095.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of EMNLP*, pages 469–478.

David Kamholz, Jonathan Pool, and Susan M. Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *Proceedings of LREC*, pages 3145–3150.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of CoNLL*, pages 211–220.

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. In *Proceedings of EMNLP*, pages 862–868.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *Proceedings of COLING*, pages 1459–1474.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit (MT SUMMIT)*, pages 79–86.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of EMNLP*, pages 5039–5049.

Robert Litschko, Goran Glavaš, Ivan Vulic, and Laura Dietz. 2019. Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of SIGIR*, pages 1109–1112.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *Proceedings of SIGIR*, pages 1253–1256.

Noa Yehezkel Lubin, Jacob Goldberger, and Yoav Goldberg. 2019. Aligning vector-spaces with noisy supervised lexicon. In *Proceedings of NAACL-HLT*, pages 460–465.

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, Gregory S Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Tasnim Mohiuddin and Shafiq Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In *Proceedings of NAACL-HLT*, pages 3857–3867.

Ndapa Nakashole. 2018. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of EMNLP*, pages 512–522.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of ACL*, pages 4990–4995.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

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, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. In *Proceedings of ACL*, pages 1713–1722.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL*, pages 778–788.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of RANLP*, pages 237–248.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*, pages 384–394.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of ACL*, pages 247–257.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of SIGIR*, pages 363–372.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of EMNLP*, pages 2536–2548.

Søren Wichmann, André Müller, Viveka Velupillai, Cecil H Brown, Eric W Holman, Pamela Brown, Sebastian Sauppe, Oleg Belyaev, Matthias Urban, Zarina Molochieva, et al. 2018. The ASJP database (version 18).

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of NAACL-HLT*, pages 1006–1011.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or shōjo? Cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of ACL*, pages 3180–3189.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – Multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of NAACL-HLT*, pages 1307–1317.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*.

Yftah Ziser and Roi Reichart. 2018. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of EMNLP*, pages 238–249.