How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions

Goran Glavaš[1], Robert Litschko[1], Sebastian Ruder[2,3]*, and Ivan Vulić[4]

[1]Data and Web Science Group, University of Mannheim, Germany
[2]Insight Research Centre, National University of Ireland, Galway, Ireland
[3]Aylien Ltd., Dublin, Ireland
[4] PolyAI Ltd., London, United Kingdom
{goran, litschko}@informatik.uni-mannheim.de,
sebastian@ruder.io, iv250@cam.ac.uk

Abstract

Cross-lingual word embeddings (CLEs) facilitate cross-lingual transfer of NLP models. Despite their ubiquitous downstream usage, increasingly popular projection-based CLE models are almost exclusively evaluated on bilingual lexicon induction (BLI). Even the BLI evaluations vary greatly, hindering our ability to correctly interpret performance and properties of different CLE models. In this work, we take the first step towards a comprehensive evaluation of CLE models: we thoroughly evaluate both supervised and unsupervised CLE models, for a large number of language pairs, on BLI and three downstream tasks, providing new insights concerning the ability of cutting-edge CLE models to support cross-lingual NLP. We empirically demonstrate that the performance of CLE models largely depends on the task at hand and that optimizing CLE models for BLI may hurt downstream performance. We indicate the most robust supervised and unsupervised CLE models and emphasize the need to reassess simple baselines, which still display competitive performance across the board. We hope our work catalyzes further research on CLE evaluation and model analysis.

1 Introduction and Motivation

Following the ubiquitous use of word embeddings in monolingual NLP tasks, research in word representation quickly broadened towards *cross-lingual word embeddings* (CLEs). CLE models learn vectors of words in two or more languages and represent them in a *shared cross-lingual word vector space* where words with similar meanings obtain similar vectors, irrespective of their language. Owing to this property, CLEs hold promise to support cross-lingual NLP by enabling multilingual modeling of meaning and facilitating cross-lingual transfer for downstream NLP tasks and under-resourced languages. CLEs are used as (cross-lingual) knowledge sources in range of tasks, such as bilingual lexicon induction (Mikolov et al., 2013), document classification (Klementiev et al., 2012), information retrieval (Vulić and Moens, 2015), dependency parsing (Guo et al., 2015), sequence labeling (Zhang et al., 2016; Mayhew et al., 2017), and machine translation (Artetxe et al., 2018c; Lample et al., 2018), among others.

Earlier work typically induces CLEs by leveraging bilingual supervision from multilingual corpora aligned at the level of sentences (Zou et al., 2013; Hermann and Blunsom, 2014; Luong et al., 2015, *inter alia*) and documents (Søgaard et al., 2015; Vulić and Moens, 2016; Levy et al., 2017, *inter alia*). A recent trend are the so-called *projection-based* CLE models[1], which post-hoc align pre-trained monolingual embeddings. Their popularity stems from competitive performance coupled with a conceptually simple design, requiring only cheap bilingual supervision (Ruder et al., 2018b): they demand word-level supervision from *seed translation dictionaries*, spanning at most several thousand word pairs (Mikolov et al., 2013; Huang et al., 2015), but it has also been shown that reliable projections can be bootstrapped from small dictionaries of 50–100 pairs (Vulić and Korhonen, 2016; Zhang et al., 2016), identical strings and cognates (Smith et al., 2017; Søgaard et al., 2018), and shared numerals (Artetxe et al., 2017).

Moreover, recent work has leveraged topological similarities between monolingual vector spaces to introduce *fully unsupervised* projection-based CLE models, not demanding any bilingual supervision (Conneau et al., 2018a; Artetxe et al., 2018b, *inter alia*). Being conceptually attractive, such weakly supervised and unsupervised CLEs have recently taken the field by storm (Grave et al., 2018; Dou

*Sebastian is now affiliated with DeepMind.

[1]In the literature the methods are sometimes referred to as mapping-based CLE approaches or offline approaches.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 710–721
Florence, Italy, July 28 - August 2, 2019.

et al., 2018; Doval et al., 2018; Hoshen and Wolf, 2018; Ruder et al., 2018a; Kim et al., 2018; Chen and Cardie, 2018; Mukherjee et al., 2018; Nakashole, 2018; Xu et al., 2018; Alaux et al., 2019).

Producing the same end result—a shared cross-lingual vector space—all CLE models are directly comparable, regardless of modelling assumptions and supervision requirements. Therefore, they can support exactly the same groups of tasks. Yet, a comprehensive evaluation of recent CLE models is missing. Limited evaluations impede comparative analyses and may lead to inadequate conclusions, as models are commonly trained to perform well on a single task. While early CLE models (Klementiev et al., 2012; Hermann and Blunsom, 2014) were evaluated on downstream tasks like text classification, a large body of recent work is judged exclusively on the task of bilingual lexicon induction (BLI). This limits our understanding of CLE methodology as: **1)** BLI is an intrinsic task, and agreement between BLI and downstream performance has been challenged (Ammar et al., 2016; Bakarov et al., 2018); **2)** BLI is not the main motivation for inducing cross-lingual embedding spaces—rather, we seek to exploit CLEs to tackle multilinguality and downstream language transfer (Ruder et al., 2018b). In other words, previous research does not evaluate the true capacity of projection-based CLE models to support cross-lingual NLP. It is unclear whether and to which extent BLI performance of (projection-based) CLE models correlates with various downstream tasks of different types.

At the moment, it is virtually impossible to directly compare all recent projection-based CLE models on BLI due to the lack of a common evaluation protocol: different papers consider different language pairs and employ different training and evaluation dictionaries. Furthermore, there is a surprising lack of testing of BLI results for statistical significance. The mismatches in evaluation yield partial conclusions and inconsistencies: on the one hand, some unsupervised models (Artetxe et al., 2018b; Hoshen and Wolf, 2018) reportedly outperform competitive supervised CLE models (Artetxe et al., 2017; Smith et al., 2017). On the other hand, the most recent supervised approaches (Doval et al., 2018; Joulin et al., 2018) report performances surpassing the best unsupervised models.

Supervised projection-based CLEs require merely small-sized translation dictionaries (up to a few thousand word pairs) and such bilingual signal is easily obtainable for most language pairs.[2] Therefore, despite the attractive zero-supervision setup, we see unsupervised CLE models practically justified only if such models can, unintuitively, indeed outperform their supervised competition.

Contributions. We provide a comprehensive comparative evaluation of a wide range of state-of-the-art—both supervised and unsupervised—projection-based CLE models. Our benchmark encompasses BLI and three cross-lingual (CL) downstream tasks of different nature: document classification (CLDC), information retrieval (CLIR), and natural language inference (XNLI). We unify evaluation protocols for all models and conduct experiments over 28 language pairs spanning diverse language types.

Besides providing a unified testbed for guiding CLE research, we aim to answer the following research questions: **1)** Is BLI performance a good predictor of downstream performance for projection-based CLE models? **2)** Can unsupervised CLE models indeed outperform their supervised counterparts? The simplest models often outperform more intricate competitors: we apply a simple bootstrapping to the basic Procrustes model (PROC-B, see §2.2) and show it is competitive across the board. We find that overfitting to BLI may severely hurt downstream performance, warranting the coupling of BLI experiments with downstream evaluations in order to paint a more informative picture of CLE models' properties.

2 Projection-Based CLEs: Methodology

In contrast to more recent unsupervised models, CLE models typically require bilingual signal: aligned words, sentences, or documents. CLE models based on sentence and document alignments have been extensively studied in previous work (Vulić and Korhonen, 2016; Upadhyay et al., 2016; Ruder et al., 2018b). Current CLE research is almost exclusively focused on projection-based CLE models; they are thus also the focus of our study.[3]

[2] We argue that, if acquiring a few thousand word translation pairs is a challenge, one probably deals with a truly under-resourced language for which it would be difficult to obtain reliable monolingual embeddings in the first place. Furthermore, there are initiatives in typological linguistics research such as the ASJP database (Wichmann et al., 2018), which offers 40-item word lists denoting the same set of concepts in all the world's languages: https://asjp.clld.org/. Indirectly, such lists can offer the initial seed supervision.

[3] These methods **a)** are not bound to any particular word embedding model (i.e., they are fully agnostic to how we

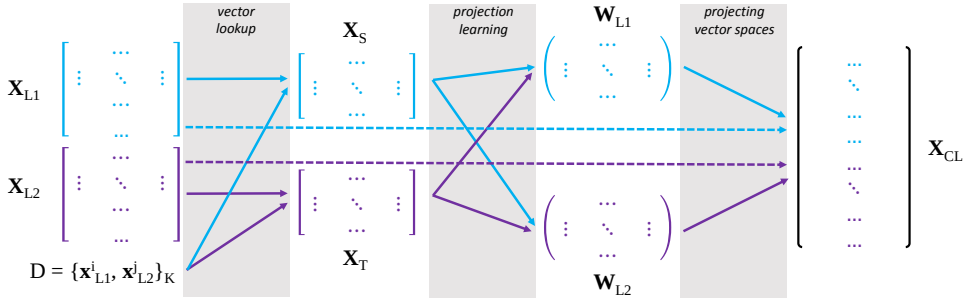

Figure 1: A general framework for post-hoc projection-based induction of cross-lingual word embeddings.

2.1 Projection-Based Framework

The goal is to learn a projection between independently trained monolingual embedding spaces. The mapping is sought using a seed bilingual lexicon, provided beforehand or extracted without supervision. A general post-hoc projection-based CLE framework is depicted in Figure 1. Let $\mathbf{X}_{L1}$ and $\mathbf{X}_{L2}$ be monolingual embedding spaces of two languages. All projection-based CLE approaches encompass the following steps:

Step 1: Construct the seed translation dictionary $D = \{(w^i_{L1}, w^j_{L2})\}^K_{k=1}$ containing K word pairs. Supervised models use an external dictionary; unsupervised models induce D automatically, assuming approximate isomorphism between $\mathbf{X}_{L1}$ and $\mathbf{X}_{L2}$.

Step 2: Align monolingual subspaces $\mathbf{X}_S = \{\mathbf{x}^i_{L1}\}^K_{k=1}$ and $\mathbf{X}_T = \{\mathbf{x}^j_{L2}\}^K_{k=1}$ using the translation dictionary D: retrieve vectors of $\{w^i_{L1}\}^K_{k=1}$ from $\mathbf{X}_{L1}$ and vectors of $\{w^j_{L2}\}^K_{k=1}$ from $\mathbf{X}_{L2}$.

Step 3: Learn to project $\mathbf{X}_{L1}$ and $\mathbf{X}_{L2}$ to the shared cross-lingual space $\mathbf{X}_{CL}$ based on the aligned matrices $\mathbf{X}_S$ and $\mathbf{X}_T$. In the general case, we learn two projection matrices $\mathbf{W}_{L1}$ and $\mathbf{W}_{L2}$: $\mathbf{X}_{CL} = \mathbf{X}_{L1}\mathbf{W}_{L1} \cup \mathbf{X}_{L2}\mathbf{W}_{L2}$. Many models, however, learn to directly project $\mathbf{X}_{L1}$ to $\mathbf{X}_{L2}$, i.e., $\mathbf{W}_{L2} = I$ and $\mathbf{X}_{CL} = \mathbf{X}_{L1}\mathbf{W}_{L1} \cup \mathbf{X}_{L2}$.

2.2 Projection-Based CLE Models

While supervised models employ external dictionaries, unsupervised models automatically induce seed translations using diverse strategies: adversarial learning (Conneau et al., 2018a), similarity-based heuristics (Artetxe et al., 2018b), PCA (Hoshen and Wolf, 2018), and optimal transport (Alvarez-Melis and Jaakkola, 2018).

Canonical Correlation Analysis (CCA). Faruqui and Dyer (2014) use CCA to project $\mathbf{X}_{L1}$ and $\mathbf{X}_{L2}$ into a shared space $\mathbf{X}_{CL}$. Projection matrices, $\mathbf{W}_{L1}$ and $\mathbf{W}_{L2}$ are obtained by applying CCA to $\mathbf{X}_S$ and $\mathbf{X}_T$. We evaluate CCA as a simple baseline that has mostly been neglected in recent BLI evaluations.

obtain monolingual vectors) and **b)** they do not require any multilingual corpora, they lend themselves to a wider spectrum of languages than the alternatives (Ruder et al., 2018b).

Algorithm 1: Bootstrapping Procrustes (PROC-B)

$\mathbf{X}_{L1}, \mathbf{X}_{L2} \leftarrow$ monolingual embeddings of $L1$ and $L2$
$D \leftarrow$ initial word translation dictionary
for *each of* n *iterations* **do**
  $\mathbf{X}_S, \mathbf{X}_T \leftarrow$ lookups for D in $\mathbf{X}_{L1}, \mathbf{X}_{L2}$
  $\mathbf{W}_{L1} \leftarrow \arg\min_W \|\mathbf{X}_S\mathbf{W} - \mathbf{X}_T\|_2$
  $\mathbf{W}_{L2} \leftarrow \arg\min_W \|\mathbf{X}_T\mathbf{W} - \mathbf{X}_S\|_2$
  $\mathbf{X'}_{L1} \leftarrow \mathbf{X}_{L1}\mathbf{W}_{L1}$; $\mathbf{X'}_{L2} \leftarrow \mathbf{X}_{L2}\mathbf{W}_{L2}$
  $D_{1,2} \leftarrow \text{nn}(\mathbf{X'}_{L1}, \mathbf{X}_{L2})$; $D_{2,1} \leftarrow \text{nn}(\mathbf{X'}_{L2}, \mathbf{X}_{L1})$
  $D \leftarrow D \cup (D_{1,2} \cap D_{2,1})$
return: $\mathbf{W}_{L1}$ (and/or $\mathbf{W}_{L2}$)

Solving the Procrustes Problem. In their seminal work, Mikolov et al. (2013) find $\mathbf{W}_{L1}$ by minimizing the Euclidean distance between projected $\mathbf{X}_S$ and $\mathbf{X}_T$: $\mathbf{W}_{L1} = \arg\min_{\mathbf{W}} \|\mathbf{X}_{L1}\mathbf{W} - \mathbf{X}_{L2}\|_2$. Xing et al. (2015) report BLI gains by imposing orthogonality on $\mathbf{W}_{L1}$. If $\mathbf{W}_{L1}$ is orthogonal, the above minimization problem becomes the Procrustes problem, with the following closed-form solution (Schönemann, 1966):

$$\mathbf{W}_{L1} = \mathbf{U}\mathbf{V}^\top, \text{ with}$$
$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = SVD(\mathbf{X}_T\mathbf{X}_S{}^\top). \quad (1)$$

The map $\mathbf{W}_{L1}$ being the solution to the Procrustes problem is the main baseline in our evaluation (PROC). Furthermore, following *self-learning* procedures that unsupervised models use to augment initially induced lexicons D (Artetxe et al., 2018b; Conneau et al., 2018a), we propose a simple bootstrapping extension of the PROC model (dubbed PROC-B). With PROC-B we aim to boost performance when starting with smaller but reliable external dictionaries. The procedure is summarized in Algorithm 1. In each iteration, we use the translation lexicon D to learn two projections: $\mathbf{W}_{L1}$ projects $\mathbf{X}_{L1}$ to $\mathbf{X}_{L2}$ and $\mathbf{W}_{L2}$ projects $\mathbf{X}_{L2}$ to $\mathbf{X}_{L1}$. Next we induce two word translations sets D_{12} and D_{21} as cross-lingual nearest neighbours

between (1) $\mathbf{X}_{L1}\mathbf{W}_{L1}$ and $\mathbf{X}_{L2}$ and (2) $\mathbf{X}_{L2}\mathbf{W}_{L2}$ and $\mathbf{X}_{L1}$. We finally augment D with mutual nearest neighbours, i.e., $D_{12} \cap D_{21}$.[4]

Discriminative Latent-Variable Model (DLV). Ruder et al. (2018a) augment the seed supervised lexicon through Expectation-Maximization in a latent-variable model. The source words $\{w^i_{L1}\}^K_{k=1}$ and target words $\{w^j_{L2}\}^K_{k=1}$ are seen as a fully connected bipartite weighted graph $G = (E, V_{L1} \cup V_{L2})$ with edges $E = V_{L1} \times V_{L2}$. By drawing embeddings from a Gaussian distribution and normalizing them, the weight of each edge $(i, j) \in E$ is shown to correspond to the cosine similarity between vectors. In the E-step, a maximal bipartite matching is found on the sparsified graph using the Jonker-Volgenant algorithm (Jonker and Volgenant, 1987). In the M-step, a better projection $\mathbf{W}_{L1}$ is learned by solving the Procrustes problem.

Ranking-Based Optimization (RCSLS). Instead of minimizing the Euclidean distance, Joulin et al. (2018) follow earlier work (Lazaridou et al., 2015) and maximize a ranking-based objective, specifically cross-domain similarity local scaling (CSLS; Conneau et al., 2018a), between the $\mathbf{X}_S\mathbf{W}_{L1}$ and $\mathbf{X}_T$. CSLS is an extension of cosine similarity commonly used for BLI inference. Let $r(\mathbf{x}^k_{L1}\mathbf{W}, \mathbf{X}_{L2})$ be the average cosine similarity of the projected source vector with its N nearest neighbors from $\mathbf{X}_{L2}$. Inversely, let $r(\mathbf{x}^k_{L2}, \mathbf{X}_{L1}\mathbf{W})$ be the average cosine similarity of a target vector with its N nearest neighbors from the projected source space $\mathbf{X}_{L1}\mathbf{W}$. By relaxing the orthogonality constraint on $\mathbf{W}_{L1}$, maximization of relaxed CSLS (dubbed RCSLS) becomes a convex optimization problem:

$$\mathbf{W}_{L1} = \arg\min_{\mathbf{W}} \frac{1}{K} \sum_{\substack{\mathbf{x}^k_{L1} \in X_S \\ \mathbf{x}^k_{L2} \in X_T}} -2\cos\left(\mathbf{x}^k_{L1}\mathbf{W}, \mathbf{x}^k_{L2}\right) + r(\mathbf{x}^k_{L1}\mathbf{W}, \mathbf{X}_{L2}) + r(\mathbf{x}^k_{L2}, \mathbf{X}_{L1}\mathbf{W}) \quad (2)$$

By maximizing (R)CSLS, this model is explicitly designed to induce cross-lingual embedding spaces that perform well in word translation, i.e., BLI.

Adversarial Alignment (MUSE). MUSE (Conneau et al., 2018a) initializes a seed bilingual lexicon solely from monolingal data using Generative Adversarial Networks (GANs; Goodfellow et al., 2014). The generator component is the linear map $\mathbf{W}_{L1}$. MUSE improves the generator $\mathbf{W}_{L1}$ by additionally competing with the discriminator (feed-forward net) that needs to distinguish between true L2 vectors from $\mathbf{X}_{L2}$ and projections from L1, $\mathbf{X}_{L1}\mathbf{W}_{L1}$. MUSE then improves the GAN-induced mapping, through a refinement step similar to the PROC-B procedure. MUSE strongly relies on the approximate isomorphism assumption (Søgaard et al., 2018), which often leads to poor GAN-based initialization, especially for distant languages.

Heuristic Alignment (VECMAP). Artetxe et al. (2018b) assume that word translations have approximately the same vectors of monolingual similarity distribution. The seed D is the set of nearest neighbors according to the similarity between monolingual similarity distribution vectors. Next, they employ a self-learning bootstrapping procedure similar to MUSE. VECMAP owes robustness to a number of empirically motivated enhancements. It adopts both multi-step pre-processing: unit length normalization, mean centering, and ZCA whitening (Bell and Sejnowski, 1997); and multiple post-processing steps: cross-correlational re-weighting, de-whitening, and dimensionality reduction (Artetxe et al., 2018a). Moreover, VECMAP critically relies on *stochastic* dictionary induction: elements of the similarity matrix are randomly set to 0 (with varying probability across iterations), allowing the model to escape poor local optima.

Iterative Closest Point Model (ICP). Hoshen and Wolf (2018) induce the seed dictionary D by projecting vectors of N most frequent words of both languages to a lower-dimensional space with PCA. They then search for an optimal alignment between L1 words (vectors $\mathbf{x}^i_1$) and L2 words (vectors $\mathbf{x}^j_2$), assuming projections $\mathbf{W}_1$ and $\mathbf{W}_2$. Let $f_1(i)$ (vice versa $f_2(j)$) be the L2 index (vice versa L1 index) to which $\mathbf{x}^i_1$ (vice versa $\mathbf{x}^j_2$) is aligned. The goal is to find projections $\mathbf{W}_1$ and $\mathbf{W}_2$ that minimize the sum of Euclidean distances between optimally-aligned vectors. Since both projections and optimal alignment are unknown, they employ a two-step Iterative Closest Point optimization algorithm that first fixes projections to find the optimal alignment and then uses that alignment to update projections, by minimizing:

$$\sum_i \|\mathbf{x}^i_1\mathbf{W}_1 - \mathbf{x}^{f_1(i)}_2\| + \sum_j \|\mathbf{x}^j_2\mathbf{W}_2 - \mathbf{x}^{f_2(j)}_1\| + \lambda \sum_i \|\mathbf{x}^i_1 - \mathbf{x}^i_1\mathbf{W}_1\mathbf{W}_2\| + \lambda \sum_j \|\mathbf{x}^j_2 - \mathbf{x}^j_2\mathbf{W}_2\mathbf{W}_1\|. \quad (3)$$

[4] We obtain best performance with just one bootstrapping iteration. Even when starting with D of size 1K, the first iteration yields between 5K and 10K mutual translations. The second iteration already produces over 20K translations, which seem to be too noisy for further performance gains.

The cyclical constraints in the second row force vectors not to change by round-projection (to the other space and back). They then employ a bootstraping procedure similar to PROC-B and MUSE and produce the final $\mathbf{W}_{L1}$ by solving Procrustes.

Gromov-Wasserstein Alignment Model (GWA). Since embedding models employ metric recovery algorithms, Alvarez-Melis and Jaakkola (2018) cast dictionary induction as optimal transport problem based on the Gromov-Wasserstein distance. They first compute intra-language costs $\mathbf{C}_{L1} = \cos(\mathbf{X}_{L1}, \mathbf{X}_{L1})$ and $\mathbf{C}_{L2} = \cos(\mathbf{X}_{L2}, \mathbf{X}_{L2})$ and inter-language similarities $\mathbf{C}_{12} = \mathbf{C}_{L1}^2\mathbf{p}\mathbb{1}_m^\top + \mathbb{1}_n\mathbf{q}(\mathbf{C}_{L2}^2)^\top$, with $\mathbf{p}$ and $\mathbf{q}$ as uniform distributions over respective vocabularies. They then induce the projections by solving the Gromov-Wasserstein optimal transport problem with a fast iterative algorithm (Peyré et al., 2016), iteratively updating parameter vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{a} = \mathbf{p} \oslash \mathbf{K}\mathbf{b}$; $\mathbf{b} = \mathbf{q} \oslash \mathbf{K}^\top\mathbf{a}$, where $\oslash$ is element-wise division and $\mathbf{K} = \exp(-\hat{\mathbf{C}}_\Gamma/\lambda)$ (with $\hat{\mathbf{C}}_\Gamma = \mathbf{C}_{12} - 2\mathbf{C}_{L1}\Gamma\mathbf{C}_{L2}^\top$). The alignment matrix $\Gamma = \text{diag}(\mathbf{a})\,\mathbf{K}\,\text{diag}(\mathbf{b})$ is recomputed in each iteration. The final projection $\mathbf{W}_{L1}$ is (again!) obtained by solving Procrustes using the final alignments from Γ as supervision.

In sum, our brief overview points to the main (dis)similarities of all projection-based CLE models: while they differ in the way the initial seed lexicon is extracted, most models are based on bootstrapping procedures that repeatedly solve the Procrustes problem from Eq. (1), typically on the trimmed vocabulary. In the final step, the fine-tuned linear map is applied on the full vocabulary.

3 Bilingual Lexicon Induction

BLI has become the *de facto* standard evaluation task for projection-based CLE models. Given a CLE space, the task is to retrieve target language translations for a (test) set of source language words. A typical BLI evaluation in the recent literature reports comparisons with the well-known MUSE model on a few language pairs, always involving English as one of the languages—a comprehensive comparative BLI evaluation conducted on a large set of language pairs is missing. Our evaluation spans 28 language pairs, many of which do not involve English.[5] Furthermore, to allow for (1) fair comparison across supervised models and (2) direct comparisons across different language pairs, we create training and evaluation dictionaries that are *fully aligned* across all evaluated language pairs. Finally, we also discuss other choices in existing BLI evaluations which are currently taken for granted: e.g., (in)appropriate evaluation metrics and lack of significance testing.

Language Pairs. Our evaluation comprises eight languages: Croatian (HR), English (EN), Finnish (FI), French (FR), German (DE), Italian (IT), Russian (RU), and Turkish (TR). For diversity, we selected two languages from three different Indo-European branches: Germanic (EN, DE), Romance (FR, IT), and Slavic (HR, RU); as well as two non-Indo-European languages (FI, TR). From these, we create a total of 28 language pairs for evaluation.

Monolingual Embeddings. Following prior work, we use 300-dimensional fastText embeddings (Bojanowski et al., 2017)[6], pretrained on complete Wikipedias of each language. We trim all vocabularies to the 200K most frequent words.

Translation Dictionaries. We automatically created translation dictionaries using Google Translate, similar to prior work (Conneau et al., 2018a). We selected the 20K most frequent English words and automatically translated them to the other seven languages. We retained only tuples for which all translations were unigrams found in vocabularies of respective monolingual embedding spaces, leaving us with ≈7K tuples. We reserved 5K tuples created from the more frequent English words for training, and the remaining 2K tuples for testing. We also created two smaller training dictionaries, by selecting tuples corresponding to 1K and 3K most frequent English words.

Evaluation Measures and Significance. BLI is generally cast as a ranking task. Existing work uses precision at rank k ($P@k$, $k \in \{1, 5, 10\}$) as a BLI evaluation metric. We advocate the use of mean average precision (MAP) instead.[7] While correlated with $P@k$, MAP is more informative: unlike MAP, $P@k$ treats all models that rank the correct translation below k equally.[8]

The limited size of BLI test sets warrants statistical significance testing. Yet, most current work

[5] English participates as one of the languages in all pairs in existing BLI evaluations, with the exception of Estonian–Finnish evaluated by Søgaard et al. (2018).

[6] https://fasttext.cc/docs/en/pretrained-vectors.html

[7] In this setup with only one correct translation for each query, MAP is equivalent to mean reciprocal rank (MRR).

[8] E.g., $P@5$ equally penalizes two models of which one ranks the translation at rank 6 and the other at rank 100K.

Supervised	Dict	All LPs	Filt. LPs	Succ. LPs
CCA	1K	.289	.404	28/28
CCA	3K	.378	.482	28/28
CCA	5K	.400	.498	28/28
PROC	1K	.299	.411	28/28
PROC	3K	.384	.487	28/28
PROC	5K	.405	.503	28/28
PROC-B	1K	.379	.485	28/28
PROC-B	3K	.398	.497	28/28
DLV	1K	.289	.400	28/28
DLV	3K	.381	.484	28/28
DLV	5K	.403	.501	28/28
RCSLS	1K	.331	.441	28/28
RCSLS	3K	.415	.511	28/28
RCSLS	5K	**.437**	**.527**	28/28
Unsupervised				
VECMAP		**.375**	**.471**	28/28
MUSE		.183	.458	13/28
ICP		.253	.424	22/28
GWA		.137	.345	15/28

Table 1: Summary of BLI performance (MAP). **All LPs**: average scores over all 28 language pairs; **Filt. LP**: average scores only over language pairs for which *all* models in evaluation yield at least one successful run; **Succ. LPs**: the number of language pairs for which we obtained at least one successful run. A run is considered *successful* if MAP ≥ 0.05.

provides no significance testing, declaring limited numeric gains (e.g., 1% $P@1$) to be relevant. We test BLI results for significance with the two-tailed t-test with Bonferroni correction (Dror et al., 2018).

Results and Discussion. Table 1 summarizes BLI performance over all 28 language pairs.[9] RCSLS (Joulin et al., 2018) displays the strongest BLI performance. This is not surprising given that its learning objective is tailored particularly for BLI. RCSLS outperforms other supervised models (CCA, PROC, and DLV) trained with exactly the same dictionaries. We confirm previous findings (Smith et al., 2017) suggesting that CCA and PROC exhibit very similar performance. We confirm that CCA and PROC, as well as DLV, are statistically indistinguishable in terms of BLI performance (even at $\alpha = 0.1$). Our bootstrapping PROC-B approach significantly ($p < 0.01$) boosts the performance of PROC when given a small translation dictionary with 1K pairs. For the same 1K dictionary, PROC-B also significantly outperforms RCSLS. Interestingly, training on 5K pairs does not significantly outperform training on 3K pairs for any of the supervised models (whereas using 3K pairs is better than using 1K). This is in line with prior findings (Vulić and Korhonen, 2016) that no significant improvements are to be expected from training linear maps on more than 5K word pairs.

The results highlight VECMAP (Artetxe et al., 2018b) as the most robust choice among unsupervised models: besides being the only model to produce successful runs for all language pairs, it also significantly outperforms other unsupervised models—both when considering all language pairs and only the subset where other models produce successful runs. However, VECMAP still performs worse ($p \leq 0.0002$) than PROC-B (trained on only 1K pairs) and all supervised models trained on 3K or 5K word pairs. Our findings challenge unintuitive claims from recent work (Artetxe et al., 2018b; Hoshen and Wolf, 2018; Alvarez-Melis and Jaakkola, 2018) that unsupervised CLE models perform on a par or even surpass supervised models.

Table 2 shows the scores for a subset of 10 language pairs using a subset of models from Table 1.[10] As expected, all models work reasonably well for major languages—this is most likely due to a higher quality of the respective monolingual embeddings, which are pre-trained on much larger corpora.[11] Language proximity also plays a critical role: on average, models achieve better BLI performance for languages from the same family (e.g., compare the results of HR–RU vs. HR–EN). The gap between the best-performing supervised model (RCSLS) and the best-performing unsupervised model (VECMAP) is more pronounced for cross-family language pairs, especially those consisting of one Germanic and one Slavic or non-Indo-European language (e.g., 19 points MAP difference for EN–RU, 14 for DE–RU and EN–FI, 10 points for EN–TR and EN–HR, 9 points for DE–FI and EN–HR). We suspect that this is due to heuristics based on intra-language similarity distributions, employed by Artetxe et al. (2018b) to induce an initial translation dictionary. This heuristic, critically relying on the approximate isomorphism assumption, is less effective the more distant the languages are.

4 Downstream Evaluation

Moving beyond the limiting BLI evaluation, we evaluate CLE models on three diverse cross-lingual downstream tasks: **1)** cross-lingual transfer for natural language inference (XNLI), a language under-

[9] We provide detailed BLI results for each of the 28 language pairs and all models in the appendix.

[10] We provide full BLI results for all 28 language pairs and all models in the appendix.

[11] For instance, EN Wikipedia is approximately 3 times larger than DE and RU Wikipedias, 19 times larger than FI Wikipedia and 46 times larger than HR Wikipedia.

Model	Dict	EN–DE	IT–FR	HR–RU	EN–HR	DE–FI	TR–FR	RU–IT	FI–HR	TR–HR	TR–RU
PROC	1K	0.458	0.615	0.269	0.225	0.264	0.215	0.360	0.187	0.148	0.168
PROC	5K	0.544	0.669	0.372	0.336	0.359	0.338	0.474	0.294	0.259	0.290
PROC-B	1K	0.521	0.665	0.348	0.296	0.354	0.305	0.466	0.263	0.210	0.230
RCSLS	1K	0.501	0.637	0.291	0.267	0.288	0.247	0.383	0.214	0.170	0.191
RCSLS	5K	0.580	0.682	0.404	0.375	0.395	0.375	0.491	0.321	0.285	0.324
VECMAP	–	0.521	0.667	0.376	0.268	0.302	0.341	0.463	0.280	0.223	0.200
Average	–	0.520	0.656	0.343	0.294	0.327	0.304	0.440	0.260	0.216	0.234

Table 2: BLI performance (MAP) with a selection of models on a subset of evaluated language pairs.

standing task; **2)** cross-lingual document classification (CLDC), a task commonly requiring shallow n-gram-level modelling, and **3)** cross-lingual information retrieval (CLIR), an unsupervised ranking task relying on coarser semantic relatedness.

4.1 Natural Language Inference

Large training corpora for NLI exist only in English (Bowman et al., 2015; Williams et al., 2018). Recently, Conneau et al. (2018b) released a multilingual XNLI corpus created by translating the development and test portions of the MultiNLI corpus (Williams et al., 2018) to 15 other languages.

Evaluation Setup. XNLI covers 5 out of 8 languages from our BLI evaluation: EN, DE, FR, RU, and TR. Our setup is straightforward: we train a well-known robust neural NLI model, Enhanced Sequential Inference Model (ESIM; Chen et al., 2017)[12] on the large English MultiNLI corpus, using EN word embeddings from a shared EN–L2 (L2 $\in$ {DE, FR, RU, TR }) embedding space. We then evaluate the model on the L2 portion of the XNLI by feeding L2 vectors from the shared space. [13]

Results and Discussion. XNLI accuracy scores are summarized in Table 3. The mismatch between BLI and XNLI performance is most obvious for RCSLS. While RCSLS is the best-performing model on BLI, it shows subpar performance on XNLI across the board. This suggests that specializing CLE spaces for word translation can seriously hurt cross-lingual transfer for language understanding tasks. As the second indication of the mismatch, the unsupervised VECMAP model, outperformed by supervised models on BLI, performs on par with PROC and PROC-B on XNLI. Finally, there

Supervised	Dict	EN–DE	EN–FR	EN–TR	EN–RU	Avg
PROC	1K	0.561	0.504	0.534	0.544	0.536
PROC	5K	0.607	0.534	0.568	0.585	**0.574**
PROC-B	1K	0.613	0.543	0.568	0.593	**0.579**
PROC-B	3K	0.615	0.532	0.573	0.599	**0.580**
DLV	5K	0.614	0.556	0.536	0.579	**0.571**
RCSLS	1K	0.376	0.357	0.387	0.378	0.374
RCSLS	5K	0.390	0.363	0.387	0.399	0.385
Unsupervised						
VECMAP		0.604	0.613	0.534	0.574	**0.581**
MUSE		0.611	0.536	0.359*	0.363*	0.467
ICP		0.580	0.510	0.400*	0.572	0.516
GWA		0.427*	0.383*	0.359*	0.376*	0.386

Table 3: XNLI performance (test set accuracy). Bold: highest scores, with mutually insignificant differences according to the non-parametric shuffling test (Yeh, 2000). Asterisks denote language pairs for which CLE models could not yield successful runs in the BLI task.

are significant differences between BLI and XNLI performance across language pairs—while we observe much better BLI performance for EN–DE and EN–FR compared to EN–RU and especially EN–TR, XNLI performance of most models for EN–RU and EN–TR surpasses that for EN–FR and is close to that for EN–DE. While this can be an artifact of the XNLI dataset creation, we support these observations for invidivual language pairs by measuring an overall Spearman correlation of only 0.13 between BLI and XNLI over individual language pairs scores (for all models).

The PROC model performs significantly better on XNLI when trained on 5K pairs than with 1K pairs, and this is consistent with BLI results. However, we show that we can reach the same performance level using 1K pairs and the proposed PROC-B bootstrapping scheme. VECMAP is again the most robust and most effective unsupervised model, but it is outperformed by the PROC-B model on more distant language pairs, EN–TR and EN–RU.

4.2 Document Classification

Evaluation Setup. Our next evaluation task is cross-lingual document classification (CLDC). We use the TED CLDC corpus (Hermann and Blun-

[12] Since our aim is to compare different bilingual spaces—input vectors for ESIM, kept fixed during training—we simply use the default ESIM hyper-parameter configuration.

[13] Our goal is not to compete with current state-of-the-art systems for (X)NLI (Artetxe and Schwenk, 2018; Lample and Conneau, 2019), but rather to provide means to analyze properties and relative performance of diverse CLE models in a downstream language understanding task.

Supervised	Dict	DE	FR	IT	RU	TR	Avg
PROC	1K	.250	.107	.158	.127	.309	.190
PROC	5K	.345	.239	.310	.251	.190	.267
PROC-B	1K	.374	.182	.205	.243	.254	.251
PROC-B	3K	.352	.210	.218	.186	.310	.255
DLV	5K	.299	.175	.234	.375	.208	.258
RCSLS	1K	.557	.550	.516	.466	.419	**.501**
RCSLS	5K	.588	.540	.451	.527	.447	**.510**
Unsupervised							
VECMAP		.433	.316	.333	.504	.439	**.405**
MUSE		.288	.223	.198	.226*	.264*	.240
ICP		.492	.254	.457	.362	.175*	.348
GWA		.180*	.209*	.206*	.151*	.173*	.184

Table 4: CLDC performance (micro-averaged F_1 scores); cross-lingual transfer EN–X. Numbers in bold denote the best scores in the model group. Asterisks denote language pairs for which CLE models did not yield successful runs in the BLI task.

som, 2014), covering 15 topics and 12 language pairs (EN is one of the languages in all pairs). A binary classifier is trained and evaluated for each topic and each language pair, using predefined train and test splits. Intersecting TED and BLI languages results in five CLDC evaluation pairs: EN–DE, EN–FR, EN–IT, EN–RU, and EN–TR. Since we seek to compare different CLEs and analyse their contribution and not to match state-of-the-art on TED, for the sake of simplicity we employ a light-weight CNN-based classifier in CLDC experiments.[14]

Results and Discussion. The CLDC results (F_1 micro-averaged over 12 classes) are shown in Table 4. In contrast to XNLI, RCSLS, the best-performing model on BLI, obtains peak scores on CLDC as well, with a wide margin w.r.t. other models. It significantly outperforms the unsupervised VECMAP model, which in turn significantly outperforms all other supervised models. Surprisingly, supervised Procrustes-based models (PROC, PROC-B, and DLV) that performed strongly on both BLI and XNLI display very weak performance on CLDC: this calls for further analyses.

4.3 Information Retrieval

Finally, we analyse behaviour of CLE models in cross-lingual information retrieval. Unlike XNLI and CLDC, we perform CLIR in an unsupervised fashion by comparing aggregate semantic representations of queries in $L1$ with aggregate semantic representations of documents in $L2$. Retrieval arguably requires more language understanding than CLDC (where we capture n-grams) and less than XNLI (modeling subtle meaning nuances).

Evaluation Setup. We employ a simple and effective unsupervised CLIR model from (Litschko et al., 2018)[15] that (1) builds query and document representations as weighted averages of word vectors from the CLE space and (2) computes the relevance score as the cosine similarity between aggregate query and document vectors. We evaluate CLE models on the standard test collections from the CLEF 2000-2003 ad-hoc retrieval Test Suite.[16] We obtain 9 language pairs by intersecting languages from CLEF and our BLI evaluation.

Results and Discussion. CLIR performance (MAP scores for all 9 language pairs) is shown in Table 5. The supervised Procrustes-based methods (PROC, PROC-B, DLV) appear to have an edge in CLIR, with the bootstrapping PROC-B model outperforming all other CLE methods.[17] Contrary to other downstream tasks, VECMAP is not the best-performing unsupervised model on CLIR—ICP performs best considering all nine language pairs (All LPs) and MUSE is best on LPs for which all models produced meaningful spaces (Filt. LPs). RCSLS, the best-performing BLI model, displays only mediocre CLIR performance.

4.4 Further Discussion

At first glance, BLI performance shows a weak and inconsistent correlation with results in downstream tasks. The behaviour of RCSLS is especially peculiar: it is the best-performing BLI model and it achieves the best results on CLDC by a wide margin, but it is not at all competitive on XNLI and falls short of other supervised models in CLIR. Downstream results of other models seem, with a few exceptions, to correspond to BLI trends.

To further investigate this, in Table 6 we measure correlations (in terms of Pearson's ρ) between aggregate task performances on BLI and each downstream task by considering (1) all models and (2) all models except RCSLS.[18] Without RCSLS,

[14] We implement a CNN with a single 1-D convolutional layer (8 filters for sizes 2–5) and a 1-max pooling layer, coupled with a softmax classifier. We minimize the negative log-likelihood using the Adam algorithm (Kingma and Ba, 2014).

[15] The model is dubbed BWE-AGG in the original paper.

[16] http://catalog.elra.info/en-us/repository/browse/ELRA-E0008/

[17] Similarly to the BLI evaluation, we test the significance by applying a Student's t-test on two lists of ranks of relevant documents (concatenated across all test collections), produced by two models under comparison. Even with Bonferroni correction for multiple tests, PROC-B significantly outperforms all other CLE models at $\alpha = 0.05$.

[18] For measuring correlation between BLI and each down-

Model	Dict	DE-FI	DE-IT	DE-RU	EN-DE	EN-FI	EN-IT	EN-RU	FI-IT	FI-RU	Avg
Supervised											
PROC	1K	0.147	0.155	0.098	0.175	0.101	0.210	0.104	0.113	0.096	0.133
PROC	5K	0.255	0.212	0.152	0.261	0.200	0.240	0.152	0.149	0.146	0.196
PROC-B	1K	0.294	0.230	0.155	0.288	0.258	0.265	0.166	0.151	0.136	**0.216**
PROC-B	3K	0.305	0.232	0.143	0.238	0.267	0.269	0.150	0.163	0.170	**0.215**
DLV	5K	0.255	0.210	0.155	0.260	0.206	0.240	0.151	0.147	0.147	0.197
RCSLS	1K	0.114	0.133	0.077	0.163	0.063	0.163	0.106	0.074	0.069	0.107
RCSLS	5K	0.196	0.189	0.122	0.237	0.127	0.210	0.133	0.130	0.113	0.162
Unupervised											
VECMAP	–	0.240	0.129	0.162	0.200	0.150	0.201	0.104	0.096	0.109	0.155
MUSE	–	0.001*	0.210	0.195	0.280	0.000*	0.272	0.002*	0.002*	0.001*	0.107
ICP	–	0.252	0.170	0.167	0.230	0.230	0.231	0.119	0.117	0.124	**0.182**
GWA	–	0.218	0.139	0.149	0.013	0.005*	0.007*	0.005*	0.058	0.052	0.072

Table 5: CLIR performance (MAP) of CLE models (the first language in each column is the query language, the second is the language of the document collection). Numbers in bold denote the best scores in the model group. Asterisks denote language pairs for which CLE models did not yield successful runs in BLI evaluation.

Models	XNLI	CLDC	CLIR
All models	0.269	0.390	0.764
All w/o RCSLS	**0.951**	0.266	**0.910**

Table 6: Correlations of model-level results between BLI and each of the three downstream tasks.

BLI results correlate strongly with XNLI and CLIR results and weakly with CLDC results.

But why does RCSLS diverge from other models? All other models induce an orthogonal projection (using given or induced dictionaries), minimizing the post-projection Euclidean distances between aligned words. In contrast, by maximizing CSLS, RCSLS relaxes the orthogonality condition imposed on the projection. This allows for distortions of the source embedding space after projection. The exact nature of these distortions and their impact on downstream performance of RCSLS require further investigation. However, these findings indicate that downstream evaluation is even more important for CLE models that learn non-orthogonal projections. For CLE models with orthogonal projections, downstream results seem to be more in line with BLI performance.

This brief task correlation analysis is based on coarse-grained model-level aggregation. The actual selection of the strongest baseline models requires finer-grained tuning at the level of particular language pairs and evaluation tasks of interest. Nonetheless, our experiments detect two robust baselines that should be included as indicative reference points in future CLE research: PROC-B (supervised) and VECMAP (unsupervised).

stream task T, we average model's BLI performance only over the language pairs included in the task T.

5 Conclusion

Rapid development of cross-lingual word embedding (CLE) methods is not met with adequate progress in their fair and systematic evaluation. CLE models are commonaly evaluated only in bilingual lexicon induction (BLI), and even the BLI task includes a variety of evaluation setups which are not directly comparable, hindering our ability to correctly interpret the key results. In this work, we have made the first step towards a comprehensive evaluation of CLEs. By systematically evaluating CLE models for many language pairs on BLI and three downstream tasks, we shed new light on the ability of current cutting-edge CLE models to support cross-lingual NLP. In particular, we have empirically proven that the quality of CLE models is largely task-dependent and that overfitting the models to the BLI task can result in deteriorated performance in downstream tasks. We have highlighted the most robust supervised and unsupervised CLE models and have exposed the need for reassessing existing baselines, as well as for unified and comprehensive evaluation protocols. We hope that this study will encourage future work on CLE evaluation and analysis and help guide the development of new CLE models. . We make the code and resources available at: https://github.com/codogogo/xling-eval.

Acknowledgments

The work of the first two authors was supported by the Baden-Württemberg Stiftung (AGREE grant of the Eliteprogramm). We thank the anonymous reviewers for their useful comments and suggestions.

References

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.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

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*.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

Amir Bakarov, Roman Suvorov, and Ilya Sochenkov. 2018. The limitations of cross-language word embeddings evaluation. In *Proceedings of *SEM*, pages 94–100.

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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of ACL*, pages 1657–1668.

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.

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. 2018. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of EMNLP*, pages 294–304.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL*, pages 1383–1392.

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

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of NIPS*, pages 2672–2680.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with Wasserstein procrustes. *CoRR*, abs/1805.11222.

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.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of ACL*, pages 58–68.

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

Kejun Huang, Matt Gardner, Evangelos Papalexakis, Christos Faloutsos, Nikos Sidiropoulos, Tom Mitchell, Partha P. Talukdar, and Xiao Fu. 2015. Translation invariant word embeddings. In *Proceedings of EMNLP*, pages 1084–1088.

Roy Jonker and Anton Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of EMNLP*, pages 2979–2984.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

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

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning. In *Proceedings of ACL-IJCNLP*, pages 270–280.

Omer Levy, Anders Søgaard, and Yoav Goldberg. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of EACL*, pages 765–774.

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.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of EMNLP*, pages 2536–2545.

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

Tanmoy Mukherjee, Makoto Yamada, and Timothy Hospedales. 2018. Learning unsupervised word translations without adversaries. In *Proceedings of EMNLP*, pages 627–632.

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

Gabriel Peyré, Marco Cuturi, and Justin Solomon. 2016. Gromov-Wasserstein averaging of kernel and distance matrices. In *Proceedings of ICML*, pages 2664–2672.

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. 2018a. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of EMNLP*, pages 458–468.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2018b. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Peter H Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10.

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.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of ACL*, pages 1661–1670.

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

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).

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

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.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of EMNLP*, pages 2465–2474.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of COLING*, pages 947–953.

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.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*, pages 1393–1398.