On the Limitations of Unsupervised Bilingual Dictionary Induction

Anders Søgaard[♡] **Sebastian Ruder**[♠♣] **Ivan Vulić**[◇]
[♡]University of Copenhagen, Copenhagen, Denmark
[♠]Insight Research Centre, National University of Ireland, Galway, Ireland
[♣]Aylien Ltd., Dublin, Ireland
[◇]Language Technology Lab, University of Cambridge, UK
soegaard@di.ku.dk,sebastian@ruder.io,iv250@cam.ac.uk

Abstract

Unsupervised machine translation—i.e., not assuming *any* cross-lingual supervision signal, whether a dictionary, translations, or comparable corpora—seems impossible, but nevertheless, Lample et al. (2018a) recently proposed a fully unsupervised machine translation (MT) model. The model relies heavily on an adversarial, unsupervised alignment of word embedding spaces for *bilingual dictionary induction* (Conneau et al., 2018), which we examine here. Our results identify the limitations of current unsupervised MT: unsupervised bilingual dictionary induction performs much worse on morphologically rich languages that are not dependent marking, when monolingual corpora from different domains or different embedding algorithms are used. We show that a simple trick, exploiting a weak supervision signal from identical words, enables more robust induction, and establish a near-perfect correlation between unsupervised bilingual dictionary induction performance and a previously unexplored graph similarity metric.

1 Introduction

Cross-lingual word representations enable us to reason about word meaning in multilingual contexts and facilitate cross-lingual transfer (Ruder et al., 2018). Early cross-lingual word embedding models relied on large amounts of parallel data (Klementiev et al., 2012; Mikolov et al., 2013a), but more recent approaches have tried to minimize the amount of supervision necessary (Vulić and Korhonen, 2016; Levy et al., 2017; Artetxe et al., 2017). Some researchers have even presented *unsupervised* methods that do not rely on any form of cross-lingual supervision at all (Barone, 2016; Conneau et al., 2018; Zhang et al., 2017).

Unsupervised cross-lingual word embeddings hold promise to induce bilingual lexicons and machine translation models in the absence of dictionaries and translations (Barone, 2016; Zhang et al., 2017; Lample et al., 2018a), and would therefore be a major step toward machine translation to, from, or even between low-resource languages.

Unsupervised approaches to learning cross-lingual word embeddings are based on the assumption that monolingual word embedding graphs are approximately isomorphic, that is, after removing a small set of vertices (words) (Mikolov et al., 2013b; Barone, 2016; Zhang et al., 2017; Conneau et al., 2018). In the words of Barone (2016):

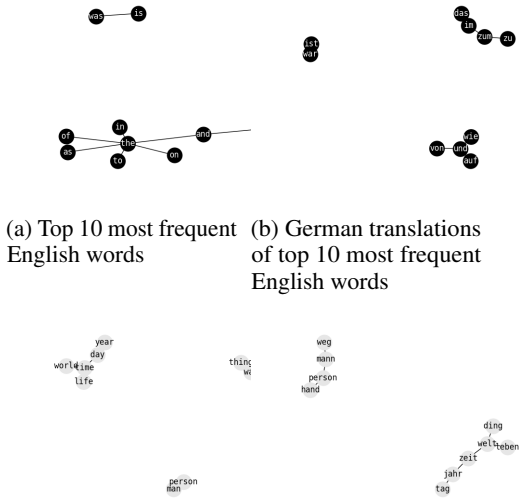
> *...we hypothesize that, if languages are used to convey thematically similar information in similar contexts, these random processes should be approximately isomorphic between languages, and that this isomorphism can be learned from the statistics of the realizations of these processes, the monolingual corpora, in principle without any form of explicit alignment.*

Our results indicate this assumption is not true in general, and that approaches based on this assumption have important limitations.

Contributions We focus on the recent state-of-the-art unsupervised model of Conneau et al. (2018).[1] Our contributions are: (a) In §2, we show that the monolingual word embeddings used in Conneau et al. (2018) are *not* approximately isomorphic, using the VF2 algorithm (Cordella et al., 2001) and we therefore introduce a metric for quantifying the similarity of word embeddings, based on Laplacian eigenvalues. (b) In §3, we identify circumstances under which the unsupervised bilingual

[1]Our motivation for this is that Artetxe et al. (2017) use small dictionary seeds for supervision, and Barone (2016) seems to obtain worse performance than Conneau et al. (2018). Our results should extend to Barone (2016) and Zhang et al. (2017), which rely on very similar methodology.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 778–788
Melbourne, Australia, July 15 - 20, 2018.



(a) Top 10 most frequent English words

(b) German translations of top 10 most frequent English words

(c) Top 10 most frequent English nouns

(d) German translations of top 10 most frequent English nouns

Figure 1: Nearest neighbor graphs.

dictionary induction (BDI) algorithm proposed in Conneau et al. (2018) does not lead to good performance. (c) We show that a simple trick, exploiting a weak supervision signal from words that are identical across languages, makes the algorithm much more robust. Our main finding is that the performance of unsupervised BDI depends heavily on all three factors: the language pair, the comparability of the monolingual corpora, and the parameters of the word embedding algorithms.

2 How similar are embeddings across languages?

As mentioned, recent work focused on unsupervised BDI assumes that monolingual word embedding spaces (or at least the subgraphs formed by the most frequent words) are approximately isomorphic. In this section, we show, by investigating the nearest neighbor graphs of word embedding spaces, that word embeddings are far from isomorphic. We therefore introduce a method for computing the similarity of non-isomorphic graphs. In §4.7, we correlate our similarity metric with performance on unsupervised BDI.

Isomorphism To motivate our study, we first establish that word embeddings are far from graph isomorphic[2]—even for two closely related languages, English and German, and using embeddings induced from comparable corpora (Wikipedia) with the same hyper-parameters.

If we take the top k most frequent words in English, and the top k most frequent words in German, and build nearest neighbor graphs for English and German using the monolingual word embeddings used in Conneau et al. (2018), the graphs are of course very different. This is, among other things, due to German case and the fact that *the* translates into *der*, *die*, and *das*, but unsupervised alignment does not have access to this kind of information. *Even* if we consider the top k most frequent English words *and their translations* into German, the nearest neighbor graphs are not isomorphic. Figure 1a-b shows the nearest neighbor graphs of the top 10 most frequent English words on Wikipedia, and their German translations.

Word embeddings are particularly good at capturing relations between nouns, but even if we consider the top k most frequent English *nouns* and their translations, the graphs are not isomorphic; see Figure 1c-d. We take this as evidence that word embeddings are not approximately isomorphic across languages. We also ran graph isomorphism checks on 10 random samples of frequent English nouns and their translations into Spanish, and only in 1/10 of the samples were the corresponding nearest neighbor graphs isomorphic.

Eigenvector similarity Since the nearest neighbor graphs are not isomorphic, even for frequent translation pairs in neighboring languages, we want to quantify the potential for unsupervised BDI using a metric that captures varying degrees of graph similarity. Eigenvalues are compact representations of global properties of graphs, and we introduce a spectral metric based on Laplacian eigenvalues (Shigehalli and Shettar, 2011) that quantifies the extent to which the nearest neighbor graphs are *isospectral*. Note that (approximately) isospectral graphs need not be (approximately) isomorphic, but (approximately) isomorphic graphs are always (approximately) isospectral (Gordon et al., 1992). Let A_1 and A_2 be the adjacency matrices of the nearest neighbor graphs G_1 and G_2 of our two word embeddings, respectively. Let $L_1 = D_1 - A_1$ and $L_2 = D_2 - A_2$ be the Laplacians of the nearest neighbor graphs, where D_1 and D_2 are the corresponding diagonal matrices of degrees. We now

[2]Two graphs that contain the same number of graph vertices connected in the same way are said to be isomorphic. In the context of weighted graphs such as word embeddings, a weak version of this is to require that the underlying nearest neighbor graphs for the most frequent k words are isomorphic.

compute the eigensimilarity of the Laplacians of the nearest neighbor graphs, L_1 and L_2. For each graph, we find the smallest k such that the sum of the k largest Laplacian eigenvalues is <90% of the Laplacian eigenvalues. We take the smallest k of the two, and use the sum of the squared differences between the largest k Laplacian eigenvalues Δ as our similarity metric.

$$\Delta = \sum_{i=1}^{k} (\lambda_{1_i} - \lambda_{2_i})^2$$

where k is chosen s.t.

$$\min_j \{ \frac{\sum_{i=1}^{k} \lambda_{j_i}}{\sum_{i=1}^{n} \lambda_{ji}} > 0.9 \}$$

Note that $\Delta = 0$ means the graphs are isospectral, and the metric goes to infinite. Thus, the higher Δ is, the *less* similar the graphs (i.e., their Laplacian spectra). We discuss the correlation between unsupervised BDI performance and approximate isospectrality or eigenvector similarity in §4.7.

3 Unsupervised cross-lingual learning

3.1 Learning scenarios

Unsupervised neural machine translation relies on BDI using cross-lingual embeddings (Lample et al., 2018a; Artetxe et al., 2018), which in turn relies on the assumption that word embedding graphs are approximately isomorphic. The work of Conneau et al. (2018), which we focus on here, also makes several implicit assumptions that may or may not be necessary to achieve such isomorphism, and which may or may not scale to low-resource languages. The algorithms are not intended to be limited to learning scenarios where these assumptions hold, but since they do in the reported experiments, it is important to see to what extent these assumptions are necessary for the algorithms to produce useful embeddings or dictionaries.

We focus on the work of Conneau et al. (2018), who present a fully unsupervised approach to aligning monolingual word embeddings, induced using *fastText* (Bojanowski et al., 2017). We describe the learning algorithm in §3.2. Conneau et al. (2018) consider a specific set of learning scenarios:

(a) The authors work with the following **languages**: English-{French, German, Chinese, Russian, Spanish}. These languages, except French, are dependent marking (Table 1).[3] We evaluate Conneau et al. (2018) on (English to) Estonian (ET), Finnish (FI), Greek (EL), Hungarian (HU), Polish (PL), and Turkish (TR) in §4.2, to test whether the selection of languages in the original study introduces a bias.

(b) The monolingual corpora in their experiments are comparable; Wikipedia corpora are used, except for an experiment in which they include Google Gigawords. We evaluate across different **domains**, i.e., on all combinations of Wikipedia, EuroParl, and the EMEA medical corpus, in §4.3. We believe such scenarios are more realistic for low-resource languages.

(c) The monolingual embedding models are induced using the same **algorithms** with the same **hyper-parameters**. We evaluate Conneau et al. (2018) on pairs of embeddings induced with different hyper-parameters in §4.4. While keeping hyper-parameters fixed is always possible, it is of practical interest to know whether the unsupervised methods work on any set of pre-trained word embeddings.

We also investigate the sensitivity of unsupervised BDI to the **dimensionality** of the monolingual word embeddings in §4.5. The motivation for this is that dimensionality reduction will alter the geometric shape and remove characteristics of the embedding graphs that are important for alignment; but on the other hand, lower dimensionality introduces regularization, which will make the graphs more similar. Finally, in §4.6, we investigate the impact of different types of query **test words** on performance, including how performance varies across part-of-speech word classes and on shared vocabulary items.

3.2 Summary of Conneau et al. (2018)

We now introduce the method of Conneau et al. (2018).[4] The approach builds on existing work on learning a mapping between monolingual word embeddings (Mikolov et al., 2013b; Xing et al., 2015) and consists of the following steps: 1) **Monolingual word embeddings:** An off-the-shelf word embedding algorithm (Bojanowski et al., 2017) is used to learn source and target language spaces X

[3] A dependent-marking language marks agreement and case more commonly on dependents than on heads.

[4] https://github.com/facebookresearch/MUSE

and Y. 2) **Adversarial mapping:** A translation matrix W is learned between the spaces X and Y using adversarial techniques (Ganin et al., 2016). A discriminator is trained to discriminate samples from the translated source space WX from the target space Y, while W is trained to prevent this. This, again, is motivated by the assumption that source and target language word embeddings are approximately isomorphic. 3) **Refinement (Procrustes analysis):** W is used to build a small bilingual dictionary of frequent words, which is pruned such that only bidirectional translations are kept (Vulić and Korhonen, 2016). A new translation matrix W that translates between the spaces X and Y of these frequent word pairs is then induced by solving the Orthogonal Procrustes problem:

$$W^* = \text{argmin}_W \|WX - Y\|_F = UV^\top \\ \text{s.t. } U\Sigma V^\top = \text{SVD}(YX^\top) \quad (1)$$

This step can be used iteratively by using the new matrix W to create new seed translation pairs. It requires frequent words to serve as reliable anchors for learning a translation matrix. In the experiments in Conneau et al. (2018), as well as in ours, the iterative Procrustes refinement improves performance across the board. 4) **Cross-domain similarity local scaling** (CSLS) is used to expand high-density areas and condense low-density ones, for more accurate nearest neighbor calculation, CSLS reduces the hubness problem in high-dimensional spaces (Radovanović et al., 2010; Dinu et al., 2015). It relies on the mean similarity of a source language embedding x to its K target language nearest neighbours ($K = 10$ suggested) $nn_1, \ldots, nn_K$:

$$mnn_T(x) = \frac{1}{K}\sum_{i=1}^{K} cos(x, nn_i) \quad (2)$$

where cos is the cosine similarity. $mnn_S(y)$ is defined in an analogous manner for any target language embedding y. $CSLS(x, y)$ is then calculated as follows:

$$2cos(x, y) - mnn_T(x) - mnn_S(y) \quad (3)$$

3.3 A simple supervised method

Instead of learning cross-lingual embeddings completely without supervision, we can extract inexpensive supervision signals by harvesting identically spelled words as in, e.g. (Artetxe et al., 2017; Smith et al., 2017). Specifically, we use identically spelled words that occur in the vocabularies of both languages as bilingual seeds, without employing any additional transliteration or lemmatization/normalization methods. Using this seed dictionary, we then run the refinement step using Procrustes analysis of Conneau et al. (2018).

4 Experiments

In the following experiments, we investigate the robustness of unsupervised cross-lingual word embedding learning, varying the language pairs, monolingual corpora, hyper-parameters, etc., to obtain a better understanding of when and why unsupervised BDI works.

Task: Bilingual dictionary induction After the shared cross-lingual space is induced, given a list of N source language words $x_{u,1}, \ldots, x_{u,N}$, the task is to find a target language word t for each *query word* x_u relying on the representations in the space. t_i is the target language word closest to the source language word $x_{u,i}$ in the induced cross-lingual space, also known as the *cross-lingual nearest neighbor*. The set of learned N $(x_{u,i}, t_i)$ pairs is then run against a gold standard dictionary.

We use bilingual dictionaries compiled by Conneau et al. (2018) as gold standard, and adopt their evaluation procedure: each test set in each language consists of 1500 gold translation pairs. We rely on CSLS for retrieving the nearest neighbors, as it consistently outperformed the cosine similarity in all our experiments. Following a standard evaluation practice (Vulić and Moens, 2013; Mikolov et al., 2013b; Conneau et al., 2018), we report *Precision at 1* scores (P@1): how many times one of the correct translations of a source word w is retrieved as the nearest neighbor of w in the target language.

4.1 Experimental setup

Our default experimental setup closely follows the setup of Conneau et al. (2018). For each language we induce monolingual word embeddings for all languages from their respective tokenized and lowercased Polyglot Wikipedias (Al-Rfou et al., 2013) using *fastText* (Bojanowski et al., 2017). Only words with more than 5 occurrences are retained for training. Our *fastText* setup relies on skip-gram with negative sampling (Mikolov et al., 2013a) with standard hyper-parameters: bag-of-words contexts with the window size 2, 15 negative samples, subsampling rate 10^{-4}, and character n-gram length

	Marking	Type	# Cases
English (EN)	dependent	isolating	None
French (FR)	mixed	fusional	None
German (DE)	dependent	fusional	4
Chinese (ZH)	dependent	isolating	None
Russian (RU)	dependent	fusional	6–7
Spanish (ES)	dependent	fusional	None
Estonian (ET)	mixed	agglutinative	10+
Finnish (FI)	mixed	agglutinative	10+
Greek (EL)	double	fusional	3
Hungarian (HU)	dependent	agglutinative	10+
Polish (PL)	dependent	fusional	6–7
Turkish (TR)	dependent	agglutinative	6–7

Table 1: Languages in Conneau et al. (2018) and in our experiments (lower half)

	Unsupervised (Adversarial)	**Supervised** (Identical)	**Similarity** (Eigenvectors)
EN-ES	81.89	**82.62**	2.07
EN-ET	00.00	**31.45**	6.61
EN-FI	00.09	**28.01**	7.33
EN-EL	00.07	**42.96**	5.01
EN-HU	45.06	**46.56**	3.27
EN-PL	46.83	**52.63**	2.56
EN-TR	32.71	**39.22**	3.14
ET-FI	**29.62**	24.35	3.98

Table 2: Bilingual dictionary induction scores (P@1×100%) using **a)** the unsupervised method with adversarial training; **b)** the supervised method with a bilingual seed dictionary consisting of identical words (shared between the two languages). The third columns lists eigenvector similarities between 10 randomly sampled source language nearest neighbor subgraphs of 10 nodes and the subgraphs of their translations, all from the benchmark dictionaries in Conneau et al. (2018).

3-6. All embeddings are 300-dimensional.

As we analyze the impact of various modeling assumptions in the following sections (e.g., domain differences, algorithm choices, hyper-parameters), we also train monolingual word embeddings using other corpora and different hyper-parameter choices. Quick summaries of each experimental setup are provided in the respective subsections.

4.2 Impact of language similarity

Conneau et al. (2018) present results for several target languages: Spanish, French, German, Russian, Chinese, and Esperanto. All languages but Esperanto are isolating or exclusively concatenating languages from a morphological point of view. All languages but French are dependent-marking. Table 1 lists three important morphological properties of the languages involved in their/our experiments.

Agglutinative languages with mixed or double marking show more morphological variance with content words, and we speculate whether unsupervised BDI is challenged by this kind of morphological complexity. To evaluate this, we experiment with Estonian and Finnish, and we include Greek, Hungarian, Polish, and Turkish to see how their approach fares on combinations of these two morphological traits.

We show results in the left column of Table 2. The results are quite dramatic. The approach achieves impressive performance for Spanish, one of the languages Conneau et al. (2018) include in their paper. For the languages we add here, performance is less impressive. For the languages with dependent marking (Hungarian, Polish, and Turkish), P@1 scores are still reasonable, with Turkish being slightly lower (0.327) than the others. However, for Estonian and Finnish, the method fails completely. Only in less than 1/1000 cases does a nearest neighbor search in the induced embeddings return a correct translation of a query word.[5]

The sizes of Wikipedias naturally vary across languages: e.g., *fastText* trains on approximately 16M sentences and 363M word tokens for Spanish, while it trains on 1M sentences and 12M words for Finnish. However, the difference in performance cannot be explained by the difference in training data sizes. To verify that near-zero performance in Finnish is not a result of insufficient training data, we have conducted another experiment using the large **Finnish WaC corpus** (Ljubešić et al., 2016) containing 1.7B words in total (this is similar in size to the English Polyglot Wikipedia). However, even with this large Finnish corpus, the model does not induce anything useful: P@1 equals 0.0.

We note that while languages with mixed marking may be harder to align, it seems unsupervised BDI is possible between similar, mixed marking languages. So while unsupervised learning fails for English-Finnish and English-Estonian, performance is reasonable and stable for the more similar Estonian-Finnish pair (Table 2). In general, unsupervised BDI, using the approach in Conneau et al. (2018), seems challenged when pairing En-

[5]We note, though, that varying our random seed, performance for Estonian, Finnish, and Greek is sometimes (approximately 1 out of 10 runs) *on par* with Turkish. Detecting main causes and remedies for the inherent instability of adversarial training is one the most important avenues for future research.

glish with languages that are not isolating and do not have dependent marking.[6]

The promise of zero-supervision models is that we can learn cross-lingual embeddings even for low-resource languages. On the other hand, a similar distribution of embeddings requires languages to be similar. This raises the question whether we need fully unsupervised methods at all. In fact, our supervised method that relies on very naive supervision in the form of identically spelled words leads to competitive performance for similar language pairs and better results for dissimilar pairs. The fact that we can reach competitive and more robust performance with such a simple heuristic questions the true applicability of fully unsupervised approaches and suggests that it might often be better to rely on available weak supervision.

4.3 Impact of domain differences

Monolingual word embeddings used in Conneau et al. (2018) are induced from Wikipedia, a near-parallel corpus. In order to assess the sensitivity of unsupervised BDI to the comparability and domain similarity of the monolingual corpora, we replicate the experiments in Conneau et al. (2018) using combinations of word embeddings extracted from three different domains: **1)** parliamentary proceedings from EuroParl.v7 (Koehn, 2005), **2)** Wikipedia (Al-Rfou et al., 2013), and **3)** the EMEA corpus in the medical domain (Tiedemann, 2009). We report experiments with three language pairs: English-{Spanish, Finnish, Hungarian}.

To control for the corpus size, we restrict each corpus in each language to 1.1M sentences in total (i.e., the number of sentences in the smallest, EMEA corpus). 300-dim *fastText* vectors are induced as in §4.1, retaining all words with more than 5 occurrences in the training data. For each pair of monolingual corpora, we compute their domain (dis)similarity by calculating the Jensen-Shannon divergence (El-Gamal, 1991), based on term distributions.[7] The domain similarities are displayed in Figures 2a–c.[8]

We show the results of unsupervised BDI in Figures 2g–i. For Spanish, we see good performance in all three cases where the English and Spanish corpora are from the same domain. *When the two corpora are from different domains, performance is close to zero.* For Finnish and Hungarian, performance is always poor, suggesting that more data is needed, even when domains are similar. This is in sharp contrast with the results of our minimally supervised approach (Figures 2d–f) based on identical words, which achieves decent performance in many set-ups.

We also observe a strong decrease in P@1 for English-Spanish (from 81.19% to 46.52%) when using the smaller Wikipedia corpora. This result indicates the importance of procuring large monolingual corpora from similar domains in order to enable unsupervised dictionary induction. However, resource-lean languages, for which the unsupervised method was designed in the first place, cannot be guaranteed to have as large monolingual training corpora as available for English, Spanish or other major resource-rich languages.

4.4 Impact of hyper-parameters

Conneau et al. (2018) use the same hyper-parameters for inducing embeddings for all languages. This is of course always practically possible, but we are interested in seeing whether their approach works on pre-trained embeddings induced with possibly very different hyper-parameters. We focus on two hyper-parameters: context window-size (*win*) and the parameter controlling the number of n-gram features in the *fastText* model (*chn*), while at the same time varying the underlying algorithm: *skip-gram* vs. *cbow*. The results for English-Spanish are listed in Table 3.

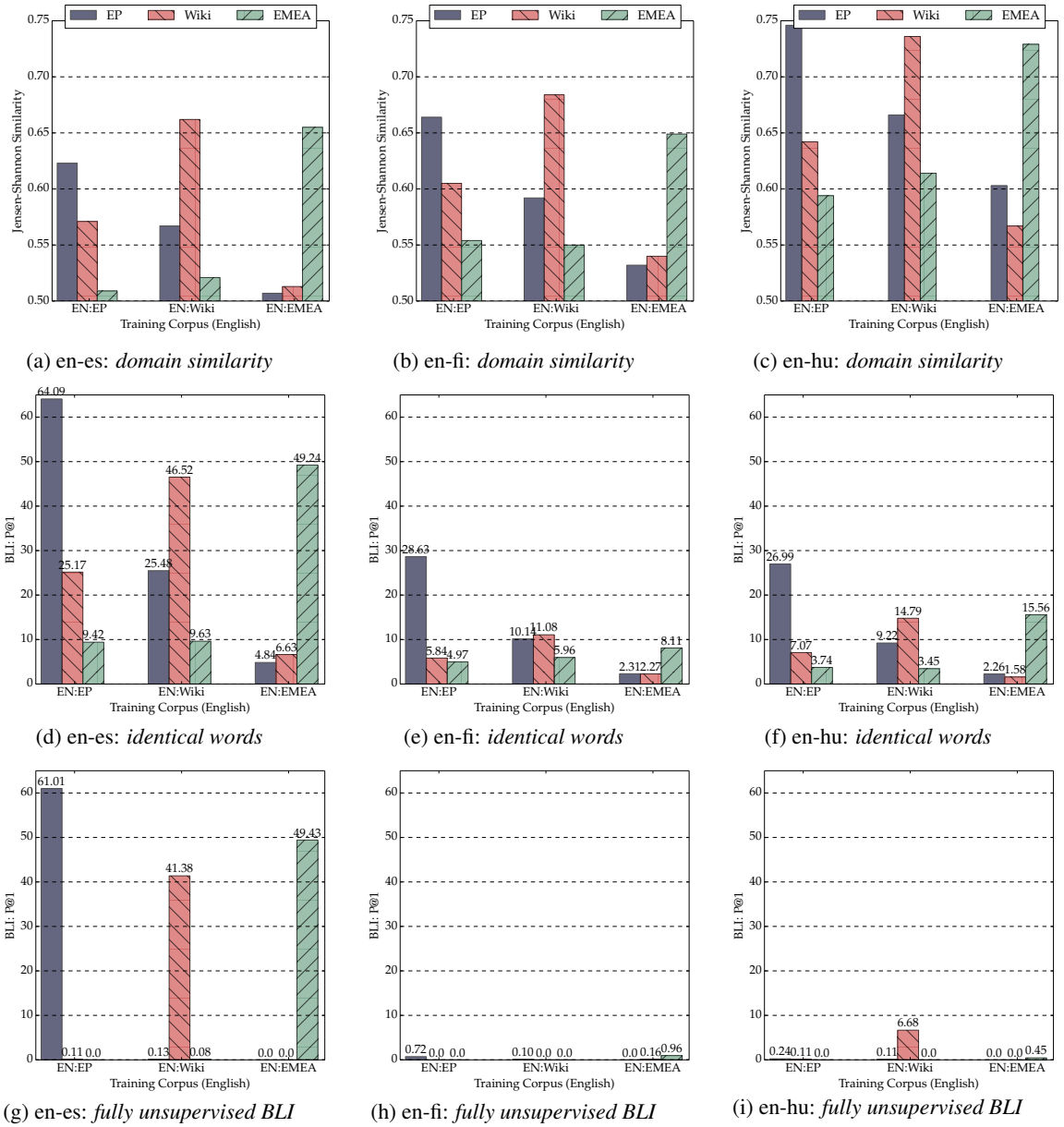
The small variations in the hyper-parameters with the same underlying algorithm (i.e., using *skip-gram* or *cbow* for both EN and ES) yield only slight drops in the final scores. Still, the best scores are obtained with the same configuration on both sides. Our main finding here is that unsupervised BDI fails (even) for EN-ES when the two monolingual embedding spaces are induced by two different algorithms (see the results of the entire Spanish *cbow* column).[9] In sum, this means that *the unsupervised approach is unlikely to work on pre-trained word embeddings unless they are induced on same-*

[6] One exception here is French, which they include in their paper, but French arguably has a relatively simple morphology.

[7] In order to get comparable term distributions, we translate the source language to the target language using the bilingual dictionaries provided by Conneau et al. (2018).

[8] We also computed $\mathcal{A}$-distances (Blitzer et al., 2007) and confirmed that trends were similar.

[9] We also checked if this result might be due to a lower-quality monolingual ES space. However, monolingual word similarity scores on available datasets in Spanish show performance comparable to that of Spanish *skip-gram* vectors: e.g., Spearman's ρ correlation is ≈ 0.7 on the ES evaluation set from SemEval-2017 Task 2 (Camacho-Collados et al., 2017).



(a) en-es: *domain similarity* (b) en-fi: *domain similarity* (c) en-hu: *domain similarity*

(d) en-es: *identical words* (e) en-fi: *identical words* (f) en-hu: *identical words*

(g) en-es: *fully unsupervised BLI* (h) en-fi: *fully unsupervised BLI* (i) en-hu: *fully unsupervised BLI*

Figure 2: Influence of language-pair *and* domain similarity on BLI performance, with three language pairs (en-es/fi/hu). **Top row, (a)-(c)**: Domain similarity (higher is more similar) computed as $dsim = 1 - JS$, where JS is Jensen-Shannon divergence; **Middle row, (d)-(f)**: baseline BLI model which learns a linear mapping between two monolingual spaces based on a set of identical (i.e., shared) words; **Bottom row, (g)-(i)**: fully unsupervised BLI model relying on the distribution-level alignment and adversarial training. Both BLI models apply the Procrustes analysis and use CSLS to retrieve nearest neighbours.

or comparable-domain, reasonably-sized training data using the same underlying algorithm.

4.5 Impact of dimensionality

We also perform an experiment on 40-dimensional monolingual word embeddings. This leads to reduced expressivity, and can potentially make the geometric shapes of embedding spaces harder to align; on the other hand, reduced dimensionality may also lead to less overfitting. We generally see worse performance (P@1 is 50.33 for Spanish, 21.81 for Hungarian, 20.11 for Polish, and 22.03 for Turkish) – but, very *interestingly*, we obtain *better performance* for Estonian (13.53), Finnish (15.33), and Greek (24.17) than we did with 300 dimensions. We hypothesize this indicates monolingual word embedding algorithms over-fit to some of the rarer peculiarities of these languages.

	English (skipgram, win=2, chn=3-6)	
	Spanish *(skipgram)*	**Spanish** *(cbow)*
==	81.89	00.00
$\neq$ win=10	81.28	00.07
$\neq$ chn=2-7	80.74	00.00
$\neq$ win=10, chn=2-7	80.15	00.13

Table 3: Varying the underlying *fastText* algorithm and hyper-parameters. The first column lists differences in training configurations between English and Spanish monolingual embeddings.

	en-es	en-hu	en-fi
Noun	80.94	26.87	00.00
Verb	66.05	25.44	00.00
Adjective	85.53	53.28	00.00
Adverb	80.00	51.57	00.00
Other	73.00	53.40	00.00

Table 4: P@1 $\times$ 100% scores for query words with different parts-of-speech.

4.6 Impact of evaluation procedure

BDI models are evaluated on a held-out set of query words. Here, we analyze the performance of the unsupervised approach across different parts-of-speech, frequency bins, and with respect to query words that have orthographically identical counterparts in the target language with the same or a different meaning.

Part-of-speech We show the impact of the part-of-speech of the query words in Table 4; again on a representative subset of our languages. The results indicate that performance on verbs is lowest across the board. This is consistent with research on distributional semantics and verb meaning (Schwartz et al., 2015; Gerz et al., 2016).

Frequency We also investigate the impact of the frequency of query words. We calculate the word frequency of English words based on Google's Trillion Word Corpus: query words are divided in groups based on their rank – i.e., the first group contains the top 100 most frequent words, the second one contains the 101th-1000th most frequent words, etc. – and plot performance (P@1) relative to rank in Figure 3. For EN-FI, P@1 was 0 across all frequency ranks. The plot shows sensitivity to frequency for HU, but less so for ES.

Homographs Since we use identical word forms (homographs) for supervision, we investigated

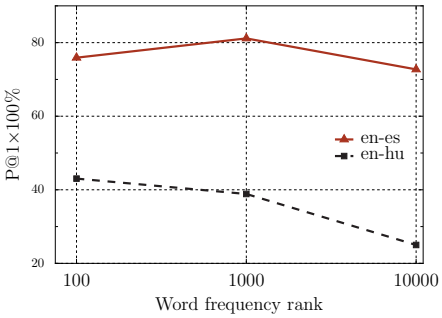


Figure 3: P@1 scores for EN-ES and EN-HU for queries with different frequency ranks.

Spelling	Meaning	en-es	en-hu	en-fi
Same	Same	45.94	18.07	00.00
Same	Diff	39.66	29.97	00.00
Diff	Diff	62.42	34.45	00.00

Table 5: Scores (P@1 $\times$ 100%) for query words with same and different spellings and meanings.

whether these are representative or harder to align than other words. Table 5 lists performance for three sets of query words: (a) source words that have homographs (words that are spelled the same way) with the same meaning (homonyms) in the target language, e.g., many proper names; (b) source words that have homographs that are not homonyms in the target language, e.g., many short words; and (c) other words. Somewhat surprisingly, words which have translations that are homographs, are associated with *lower* precision than other words. This is probably due to loan words and proper names, but note that using homographs as supervision for alignment, we achieve high precision for this part of the vocabulary *for free*.

4.7 Evaluating eigenvector similarity

Finally, in order to get a better understanding of the limitations of unsupervised BDI, we correlate the graph similarity metric described in §2 (right column of Table 2) with performance across languages (left column). Since we already established that the monolingual word embeddings are far from isomorphic—in contrast with the intuitions motivating previous work (Mikolov et al., 2013b; Barone, 2016; Zhang et al., 2017; Conneau et al., 2018)—we would like to establish another diagnostic metric that identifies embedding spaces for which the approach in Conneau et al. (2018) is likely to work. Differences in morphology, domain, or embedding parameters seem to be predictive of poor performance, but a metric that is independent of linguistic

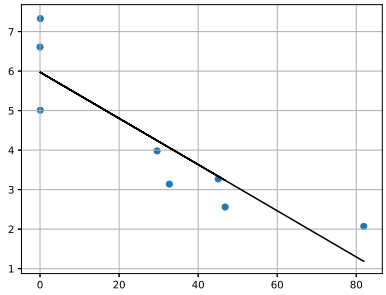


Figure 4: Strong correlation ($\rho = 0.89$) between BDI performance (x) and graph similarity (y)

categorizations and the characteristics of the monolingual corpora would be more widely applicable. We plot the values in Table 2 in Figure 4. Recall that our graph similarity metric returns a value in the half-open interval $[0, \infty)$. The correlation between BDI performance and graph similarity is strong ($\rho \sim 0.89$).

5 Related work

Cross-lingual word embeddings Cross-lingual word embedding models typically, unlike Conneau et al. (2018), require aligned words, sentences, or documents (Levy et al., 2017). Most approaches based on word alignments learn an explicit mapping between the two embedding spaces (Mikolov et al., 2013b; Xing et al., 2015). Recent approaches try to minimize the amount of supervision needed (Vulić and Korhonen, 2016; Artetxe et al., 2017; Smith et al., 2017). See Upadhyay et al. (2016) and Ruder et al. (2018) for surveys.

Unsupervised cross-lingual learning Haghighi et al. (2008) were first to explore unsupervised BDI, using features such as context counts and orthographic substrings, and canonical correlation analysis. Recent approaches use adversarial learning (Goodfellow et al., 2014) and employ a discriminator, trained to distinguish between the translated source and the target language space, and a generator learning a translation matrix (Barone, 2016). Zhang et al. (2017), in addition, use different forms of regularization for convergence, while Conneau et al. (2018) uses additional steps to refine the induced embedding space.

Unsupervised machine translation Research on unsupervised machine translation (Lample et al., 2018a; Artetxe et al., 2018; Lample et al., 2018b) has generated a lot of interest recently with a promise to support the construction of MT systems for and between resource-poor languages. All unsupervised NMT methods critically rely on accurate unsupervised BDI and back-translation. Models are trained to reconstruct a corrupted version of the source sentence and to translate its translated version back to the source language. Since the crucial input to these systems are indeed cross-lingual word embedding spaces induced in an unsupervised fashion, in this paper we also implicitly investigate one core limitation of such unsupervised MT techniques.

6 Conclusion

We investigated when unsupervised BDI (Conneau et al., 2018) is possible and found that differences in morphology, domains or word embedding algorithms may challenge this approach. Further, we found eigenvector similarity of sampled nearest neighbor subgraphs to be predictive of unsupervised BDI performance. We hope that this work will guide further developments in this new and exciting field.

Acknowledgments

We thank the anonymous reviewers, as well as Hinrich Schütze and Yova Kementchedjhieva, for their valuable feedback. Anders is supported by the ERC Starting Grant LOWLANDS No. 313695 and a Google Focused Research Award. Sebastian is supported by Irish Research Council Grant Number EBPPG/2014/30 and Science Foundation Ireland Grant Number SFI/12/RC/2289. Ivan is supported by the ERC Consolidator Grant LEXICAL No. 648909.

References

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

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, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of ICLR (Conference Track)*.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel

text trained with adversarial autoencoders. *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*, 1, pages 440–447.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:125–136.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of SEMEVAL*, pages 15–26.

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

L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. 2001. An improved algorithm for matching large graphs. *Proceedings of the 3rd IAPR TC-15 Workshop on Graphbased Representations in Pattern Recognition*, 17:1–35.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of ICLR (Workshop Papers)*.

M. A El-Gamal. 1991. The role of priors in active Bayesian learning in the sequential statistical decision framework. In *Maximum Entropy and Bayesian Methods*, pages 33–38. Springer Netherlands.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182.

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

Carolyn Gordon, David L. Webb, and Scott Wolpert. 1992. One cannot hear the shape of a drum. *Bulletin of the American Mathematical Society*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL*, pages 771–779.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *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, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR (Conference Papers)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. *CoRR*, abs/1804.07755.

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.

Nikola Ljubešić, Tommi Pirinen, and Antonio Toral. 2016. Finnish Web corpus fiWaC 1.0. Slovenian language resource repository CLARIN.SI.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation.

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2018. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267.

Vijayalaxmi Shigehalli and Vidya Shettar. 2011. Spectral technique using normalized adjacency matrices for graph matching. *International Journal of Computational Science and Mathematics*, 3:371–378.

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 (Conference Papers)*.

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

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. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of EMNLP*, pages 1613–1624.

Chao Xing, Chao Liu, Dong Wang, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of NAACL-HLT*, pages 1005–1010.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of ACL*, pages 1959–1970.