Improving Bilingual Lexicon Induction with Unsupervised Post-Processing of Monolingual Word Vector Spaces

Ivan Vulić[◇] **Anna Korhonen**[◇] **Goran Glavaš**[♣]
[◇] Language Technology Lab, TAL, University of Cambridge
[♣] Data and Web Science Group, University of Mannheim
{iv250,alk23}@cam.ac.uk goran@informatik.uni-mannheim.de

Abstract

Work on projection-based induction of cross-lingual word embedding spaces (CLWEs) predominantly focuses on the improvement of the projection (i.e., mapping) mechanisms. In this work, in contrast, we show that a simple method for post-processing monolingual embedding spaces facilitates learning of the cross-lingual alignment and, in turn, substantially improves bilingual lexicon induction (BLI). The post-processing method we examine is grounded in the generalisation of first- and second-order monolingual similarities to the n^{th}-order similarity. By post-processing monolingual spaces before the cross-lingual alignment, the method can be coupled with any projection-based method for inducing CLWE spaces. We demonstrate the effectiveness of this simple monolingual post-processing across a set of 15 typologically diverse languages (i.e., 15×14 BLI setups), and in combination with two different projection methods.

1 Introduction

Cross-lingual word embeddings (CLWEs) are a mainstay of modern cross-lingual NLP (Ruder et al., 2019b). CLWE models induce a *shared cross-lingual vector space* in which words with similar meanings obtain similar vectors regardless of their language. Their usefulness has been attested in tasks such as bilingual lexicon induction (BLI) (Gouws et al., 2015; Heyman et al., 2017), information retrieval (Litschko et al., 2018), machine translation (Artetxe et al., 2018b; Lample et al., 2018), document classification (Klementiev et al., 2012), and many others (Ruder et al., 2019b).

Importantly, CLWEs are one of the central mechanisms for facilitating transfer of language technologies for low-resource languages, which often lack sufficient bilingual signal for obvious transfer via machine translation. Lack of language resources is the main reason for popularity of the so-called *projection-based* CLWE methods (Mikolov et al., 2013a; Artetxe et al., 2016, 2018a). These models align two independently trained monolingual word vector spaces post-hoc, using limited bilingual supervision in the form of several hundred to several thousand word translation pairs (Mikolov et al., 2013a; Vulić and Korhonen, 2016; Joulin et al., 2018; Ruder et al., 2018). Some models even align the monolingual spaces using only identical strings (Smith et al., 2017; Søgaard et al., 2018) or numerals (Artetxe et al., 2017). The most recent work focused on fully unsupervised CLWE induction: they extract seed translation lexicons relying on topological similarities between monolingual spaces (Conneau et al., 2018; Artetxe et al., 2018a; Hoshen and Wolf, 2018; Alaux et al., 2019).

In this work, we do not focus on projection itself: rather, we investigate a transformation of *input monolingual word vector spaces* that facilitates the projection and leads to higher quality CLWEs. Regardless of the actual projection method, the quality of the input monolingual spaces has a profound impact on the induced shared cross-lingual space, and, in turn, on the quality of induced bilingual lexicons. We demonstrate that simple *unsupervised post-processing* of monolingual embedding spaces leads to substantial BLI performance gains across a large number of language pairs. Our work is inspired by observations that monolingual *"embeddings capture more information than what is immediately obvious"* (Artetxe et al., 2018c). In other words, the information surfaced in the pretrained monolingual vector spaces may not be optimal for an application such as word-level translation (BLI).

We rely on a monolingual post-processing method of Artetxe et al. (2018c): a linear transformation controlled by a single parameter that adjusts the similarity order of the input embedding spaces. We demonstrate that applying this trans-

Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP-2020), pages 45–54
July 9, 2020.

formation on both monolingual spaces before any standard projection-based CLWE framework yields consistent BLI gains for a wide array of languages. We run a large-scale BLI evaluation with 15 typologically diverse languages (i.e., $15\times14 = 210$ BLI setups) and show that this simple monolingual post-processing yields gains in 183/210 setups over the current state-of-the-art BLI models which combine self-learning (Artetxe et al., 2018a) with (weak) word-level supervision (Vulić et al., 2019). We further show that this monolingual post-processing yields improvements on other BLI datasets (Glavaš et al., 2019), for different projection-based CLWE models, and also for BLI with 210 similar (major European) languages (Dubossarsky et al., 2020), indicating the importance and robustness of monolingual post-processing for BLI.

2 Methodology

Projection-Based CLWEs: Preliminaries. Projection-based CLWE models learn a linear projection between two independently trained monolingual spaces – $\boldsymbol{X}$ (source language L_s) and $\boldsymbol{Z}$ (target language L_t) – using a word translation dictionary D to guide the alignment. $\boldsymbol{X}_D \subset \boldsymbol{X}$ and $\boldsymbol{Z}_D \subset \boldsymbol{Z}$ denote the row-aligned subsets of $\boldsymbol{X}$ and $\boldsymbol{Z}$ containing vectors of aligned words from D. $\boldsymbol{X}_D$ and $\boldsymbol{Z}_D$ are used to learn orthogonal projections $\boldsymbol{W}_x$ and $\boldsymbol{W}_z$ defining the bilingual space: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}_x \cup \boldsymbol{Z}\boldsymbol{W}_z$. While (weakly) supervised methods start from a readily available dictionary D, fully unsupervised models automatically induce the seed dictionary D (i.e., from monolingual data).[1]

Furthermore, it has been empirically validated (Artetxe et al., 2017; Vulić et al., 2019) that applying an *iterative self-learning* procedure leads to consistent BLI improvements, especially for distant languages and in low-data regimes. In a nutshell, at each self-learning iteration k, a dictionary $D^{(k)}$ is first used to learn the joint space $\boldsymbol{Y}^{(k)} = \boldsymbol{X}\boldsymbol{W}_x^{(k)} \cup \boldsymbol{Z}\boldsymbol{W}_z^{(k)}$. The mutual cross-lingual nearest neighbours in $\boldsymbol{Y}^{(k)}$ are then used to extract the new dictionary $D^{(k+1)}$. Relying on mutual nearest neighbours partially removes the noise, leading to better performance. For more technical details on self-learning, we refer the reader to prior work (Ruder et al., 2019a; Vulić et al., 2019).

Motivation. Most existing CLWE models ignore the properties of the initial monolingual spaces $\boldsymbol{X}$ and $\boldsymbol{Z}$ (i.e., they are taken "as-is") and focus on improving the projection. However, monolingual post-processing of $\boldsymbol{X}$ and $\boldsymbol{Z}$ prior to learning the projections may facilitate the projection and be beneficial for iterative setups such as self-learning. This intuition is already confirmed by a number of monolingual transformations, e.g., ℓ_2-normalisation, mean centering, or whitening/dewhitening, that are "by default" performed by toolkits such as MUSE (Conneau et al., 2018) and VecMap (Artetxe et al., 2018b; Zhang et al., 2019). In this work, however, we investigate a transformation to the monolingual spaces which is applied before they undergo the series of standard normalisation and centering steps.

Further, we investigate a line of research that leverages unsupervised post-processing of monolingual word vectors (Mu et al., 2018; Wang et al., 2018; Raunak et al., 2019; Tang et al., 2019) to emphasise semantic properties over syntactic aspects, typically with small gains reported on intrinsic word similarity (e.g., SimLex-999 (Hill et al., 2015)). In this work, we empirically validate that these unsupervised post-processing techniques can also be effective in cross-lingual scenarios for low-resource BLI, even when coupled with the current state-of-the-art CLWE frameworks that rely on "all the bells and whistles", such as self-learning and additional vector space preprocessing.

Unsupervised Monolingual Post-processing. We now outline the simple post-processing method of Artetxe et al. (2018c) used in this work, and then extend it to the bilingual setup. The core idea is to generalise the notion of first-and second-order similarity (Schütze, 1998)[2] to nth-order similarity. Let us define the (standard, first-order) similarity matrix of the source language space $\boldsymbol{X}$ as $\boldsymbol{M}_1(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{X}^T$ (similar for $\boldsymbol{Z}$). The second-order similarity can then be defined as $\boldsymbol{M}_2(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T$, where it holds $\boldsymbol{M}_2(\boldsymbol{X}) = \boldsymbol{M}_1(\boldsymbol{M}_1(\boldsymbol{X}))$; the nth-order similarity is then $\boldsymbol{M}_n(\boldsymbol{X}) = (\boldsymbol{X}\boldsymbol{X}^T)^n$. The embeddings of words w_i and w_j are given by the rows i and j of each $\boldsymbol{M}_n$ matrix.

We are then looking for a general linear transformation that adjusts the similarity order of input

[1] Recent empirical studies (Glavaš et al., 2019; Vulić et al., 2019) show that, under fair evaluation, (weakly) supervised methods always outperform their unsupervised counterparts. We thus base all our experiments in §4 on the weakly supervised setup; nonetheless, we observe substantial relative gains for the fully unsupervised setup as well.

[2] With second-order similarity, the similarity of two words is captured in terms of how similar they are to other words.

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

Table 1: Languages used in the main BLI experiments (Vulić et al., 2019), along with family (IE=Indo-European), morphological type, and ISO 639-1 code.

matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$. As proven by Artetxe et al. (2018c), the n^{th}-order similarity transformation can be obtained as $\boldsymbol{M}_n(\boldsymbol{X}) = \boldsymbol{M}_1(\boldsymbol{X}\boldsymbol{R}_{(n-1)/2})$, with $\boldsymbol{R}_\alpha = \boldsymbol{Q}\boldsymbol{\Delta}^\alpha$, where $\boldsymbol{Q}$ and $\boldsymbol{\Delta}$ are the matrices obtained via eigendecomposition of $\boldsymbol{X}^T\boldsymbol{X}$ ($\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Delta}\boldsymbol{Q}^T$): $\boldsymbol{\Delta}$ is a diagonal matrix containing eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$; $\boldsymbol{Q}$ is an orthogonal matrix with eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ as columns.[3]

Finally, we apply the above post-processing on both monolingual vector spaces $\boldsymbol{X}$ and $\boldsymbol{Z}$. This results in adjusted vector spaces $\boldsymbol{X}'_{\alpha_s} = \boldsymbol{X}\boldsymbol{R}_{\alpha_s}$ and $\boldsymbol{Z}'_{\alpha_t} = \boldsymbol{Z}\boldsymbol{R}_{\alpha_t}$. Transformed spaces $\boldsymbol{X}'_{\alpha_s}$ and $\boldsymbol{Z}'_{\alpha_t}$ then replace the original spaces $\boldsymbol{X}$ and $\boldsymbol{Z}$ as input to any standard projection-based CLWE method.

3 Experimental Setup

We evaluate the impact of unsupervised monolingual post-processing described in §2 on BLI, focusing on pairs of typologically diverse languages.[4] Mean reciprocal rank (MRR) is used as the main evaluation metric, reported as MRR$\times 100\%$.[5]

Training and Test Data. We exploit the training and test dictionaries compiled from PanLex (Kamholz et al., 2014) by Vulić et al. (2019): the data encompasses 15 diverse languages listed in Table 1 and a total of 210 distinct $L_s \rightarrow L_t$ BLI setups.[6] In addition, we evaluate on 15 European languages (i.e., 210 pairs) from Dubossarsky et al. (2020).[7], and on diverse language pairs from the BLI evaluation suite of Glavaš et al. (2019). Training and test dictionaries in all setups contain $5K$ and $2K$ word translation pairs, respectively. We create smaller training dictionaries (e.g., spanning 1K training translation pairs) by taking the most frequent pairs from the 5K dictionaries.

Monolingual Embeddings. We use the 300-dim vectors of Grave et al. (2018) for all languages, pretrained on Common Crawl and Wikipedia with fastText (Bojanowski et al., 2017).[8] All vocabularies are trimmed to the 200K most frequent words.

Projection-Based Framework. We base the induction of projection-based CLWEs on the well-known VecMap framework (Artetxe et al., 2018b);[9] it shows very competitive and robust BLI performance, especially for distant pairs, according to the recent comparative studies (Glavaš et al., 2019; Vulić et al., 2019; Doval et al., 2019). We analyse the impact of unsupervised monolingual postprocessing from §2 by (1) feeding the original vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ to VecMap (BASELINE), and then by (2) feeding their post-processed variants $\boldsymbol{X}'_{\alpha_s}$ and $\boldsymbol{Y}'_{\alpha_t}$ (POSTPROC). We experiment with projection model variants without and with self-learning, and with different initial dictionary sizes (5K and 1K).

Note that the POSTPROC variant requires tuning of two hyper-parameters: α_s and α_t. Due to a lack of development sets for BLI experiments, we tune the two α-parameters on a single language pair (BG–CA) via cross-validation; we grid-search over the following values: $[-0.5, -0.25, -0.15, 0, 0.15, 0.25, 0.5]$. We then keep them fixed to the following values: $\alpha_s = -0.25$, $\alpha_t = 0.15$ in all subsequent experiments.

4 Results and Discussion

Main BLI results averaged over each source language (L_s) are provided in Table 2, while additional results per language pair are available in

[3] Although the post-processing motivation stems from the desire to adjust discrete similarity orders, note that α is in fact a continuous parameter which can be carefully fine-tuned (negative values are also allowed). The code is available at: https://github.com/artetxem/uncovec.

[4] The focus of this work is on the standard BLI task; however, it has recently shown (Glavaš et al., 2019) that some downstream tasks strongly correlate with BLI.

[5] Our findings also hold for *Precision@M*, for $M \in \{1, 5\}$

[6] github.com/cambridgeltl/panlex-bli. For a detailed procedure on how the lexicons were obtained from PanLex, we refer the reader to the work of Vulić et al. (2019).

[7] The languages are English, German, Dutch, Swedish, Danish, Italian, Portuguese, Spanish, French, Romanian, Croatian, Polish, Russian, Czech, Bulgarian.

[8] Experiments with other monolingual vectors such as the original fastText and skip-gram (Mikolov et al., 2013b) trained on Wikipedia show the same trends in the final results.

[9] https://github.com/artetxem/vecmap

	BG-*	CA-*	EO-*	ET-*	EU-*	FI-*	HE-*	HU-*	ID-*	KA-*	KO-*	LT-*	NO-*	TH-*	TR-*	**Avg**
BASELINE (supervised, 5k)	34.3	33.5	30.4	30.1	22.8	32.4	28.7	35.4	26.1	25.0	23.9	30.2	33.2	15.4	28.3	28.6
BASELINE (self-learning, 5k)	36.1	35.6	33.6	31.6	24.4	34.8	29.4	37.4	27.2	26.3	25.1	31.0	35.6	14.8	29.9	30.2
POSTPROC (self-learning, 5k)	**37.6**	**36.9**	**34.8**	**33.5**	**25.7**	**37.4**	**31.2**	**39.5**	**28.1**	**28.2**	**26.6**	**33.3**	**37.3**	**15.6**	**32.3**	**31.9**
BASELINE (supervised, 1k)	14.6	12.9	9.8	11.7	6.5	11.7	9.6	14.3	8.9	7.9	6.1	11.1	12.7	4.4	9.1	10.1
BASELINE (self-learning, 1k)	34.1	32.7	30.2	29.3	21.2	32.9	26.8	35.4	24.3	23.7	20.3	28.4	33.7	10.3	27.4	27.4
POSTPROC (self-learning, 1k)	**35.3**	**34.0**	**30.6**	**31.1**	**21.3**	**35.3**	**27.9**	**37.5**	**25.1**	**25.0**	**21.3**	**30.4**	**35.1**	**11.1**	**29.8**	**28.7**
Improves for... (5k)	*13/14*	*12/14*	*13/14*	*13/14*	*10/14*	*14/14*	*11/14*	*14/14*	*11/14*	*13/14*	*12/14*	*11/14*	*14/14*	*8/14*	*14/14*	*183/210*
Improves for... (1k)	*13/14*	*13/14*	*9/14*	*13/14*	*7/14*	*14/14*	*11/14*	*14/14*	*11/14*	*12/14*	*13/14*	*13/14*	*13/14*	*11/14*	*14/14*	*181/210*

Table 2: BLI results (MRR×100%) for main models in comparison. We report the results with the supervised BASELINE model based on the VecMap framework (Artetxe et al., 2018b), *without* any self-learning (i.e., supervised only), and *with* the most robust self-learning setup according to the comparative analysis of Vulić et al. (2019). The scores are averaged over experimental setups where each of the 15 languages is used as the source language L_s (e.g., BG-* averages scores over 14 setups in which Bulgarian (BG) is the source language). $5k$ and $1k$ denote seed dictionary sizes. The **Avg** column shows averaged MRR scores for each model over all 15×14=210 BLI setups and we also report the number of BLI setups in which the POSTPROC method improves over both BASELINE models.

	RCSLS		VecMap	
Pair	BASELINE (SUP)	POSTPROC (SUP)	BASELINE (SUP+SL)	POSTPROC (SUP+SL)
DE–HR	17.2	21.2	40.9	42.5
DE–TR	21.4	23.6	38.5	39.1
FI–FR	37.8	40.3	47.5	48.9
FI–HR	18.9	23.5	38.1	39.9
HR–IT	30.2	31.4	47.8	49.1
TR–FI	23.6	26.1	37.5	39.0

Table 3: BLI scores on 6 distant language pairs from the evaluation sets of Glavaš et al. (2019). Supervised models without (SUP) and with self-learning (SUP+SL).

the supplemental material. We also observe performance gains with a "pure" supervised model variant (i.e., without self-learning), but for clarity, we focus our analysis on the more powerful baseline, with self-learning. We note improvements in 183/210 (seed dictionary size 5K) and 181/210 BLI setups (size: 1K) over the projection-based baselines that held previous peak scores using the same data (Vulić et al., 2019). This validates our intuition that monolingual vectors store more information which needs to be "uncovered" via monolingual post-processing. The effect of monolingual post-processing pertains after applying other perturbations such as ℓ_2-norm or mean centering. For some languages – e.g., FI, TR, NO – we achieve gains in all BLI setups with those languages as sources.

What is more, we have not carefully fine-tuned α_s and α_t: we note that even higher scores can be achieved by finer-grained fine-tuning in the future. For instance, setting $(\alpha_s, \alpha_t) = (-0.5, 0.25)$ instead of $(-0.25, 0.15)$ for TR–BG increases BLI score from 37.8 to 39.5; the previous peak score with BASELINE was 35.1. The baseline mapping is simply obtained by setting $(\alpha_s, \alpha_t) = (0, 0)$, and we note that the tuned post-processing validated in our work should be considered as a tunable option for any projection-based CLWE method.

We further probe the robustness of unsupervised post-processing by running experiments on additional BLI evaluation set of Glavaš et al. (2019) and with another mapping model: RCSLS (Joulin et al., 2018). While we again observe gains across a range of different model variants and with different seed dictionary sizes, we summarise a selection of results in Table 3. Finally, small but consistent improvements extend also to a set of 15 European languages from Dubossarsky et al. (2020) (see Fotnote 6): POSTPROC yields gains on average for all 15/15 source languages, and across 173/210 setups (5K seed dictionary); the global average improves from 43.9 (the strongest BASELINE) to 44.7. In summary, these results further underline the usefulness of the monolingual post-processing method.

5 Conclusion and Future Work

We have demonstrated a simple and effective method for improving bilingual lexicon induction (BLI) with projection-based cross-lingual word embeddings. The method is based on standalone unsupervised post-processing of initial monolingual word embeddings before mapping, and as such applicable to any projection-based CLWE method. We have verified the importance and robustness of this monolingual post-processing with a wide range of (dis)similar language pairs as well as in different BLI setups and with different CLWE methods.

In future work, we will test other unsupervised post-processors, and also probe similar methods that inject external lexical knowledge into monolingual word vectors towards improved BLI. We also plan to probe if similar gains still hold with recently proposed more sophisticated self-learning methods (Karan et al., 2020), non-linear mapping-based CLWE methods (Glavaš and Vulić, 2020; Mohiuddin and Joty, 2020). Another idea is to also apply a similar principle to contextualised word representations in cross-lingual settings (Schuster et al., 2019; Liu et al., 2019).

Acknowledgments

This work is supported by the ERC Consolidator Grant LEXICAL (no 648909) awarded to Anna Korhonen. Goran Glavaš is supported by the Eliteprogramm of the Baden-Württemberg Stiftung (AGREE grant). We thank the reviewers for their insightful suggestions.

References

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of EMNLP*, pages 2289–2294.

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. 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. 2018b. Unsupervised neural machine translation. In *Proceedings of ICLR*.

Mikel Artetxe, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018c. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of CoNLL*, pages 282–291.

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

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

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

Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2020. Lost in embedding space: Explaining cross-lingual task performance with eigenvalue divergence. *CoRR*, abs/2001.11136.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL*, pages 710–721.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of ACL*.

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

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

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

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

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

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.

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

Mladen Karan, Ivan Vulić, Anna Korhonen, and Goran Glavaš. 2020. Classification-based self-learning for weakly supervised bilingual lexicon induction. In *Proceedings of ACL*.

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

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.

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.

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of CoNLL*, pages 33–43.

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

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

Bari Saiful M. Mohiuddin, Tasnim and Shafiq Joty. 2020. Lnmap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. *CoRR, abs/1309.4168*.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *Proceedings of ICLR*.

Vikas Raunak, Vivek Gupta, and Florian Metze. 2019. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 235–243.

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

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019a. Unsupervised cross-lingual representation learning. In *Proceedings of ACL: Tutorial Abstracts*, pages 31–38.

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

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of NAACL-HLT*, pages 1599–1613.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

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

Shuai Tang, Mahta Mousavi, and Virginia R. de Sa. 2019. An empirical study on post-processing methods for word embeddings. *CoRR*, abs/1905.10971.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP*, pages 4406–4417.

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

Bin Wang, Fenxiao Chen, Angela Wang, and C.-C. Jay Kuo. 2018. Post-processing of word representations via variance normalization and dynamic embedding. *CoRR*, abs/1808.06305.

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

A Supplemental Material

We report main BLI results for all $15 \times 14 = 210$ language pairs based on PanLex training and test data in the supplemental material, grouped by the source language, and for two dictionary sizes: $|D| = 1,000$ and $|D| = 5,000$ (while similar relative performance is also observed with other dictionary sizes, e.g., $|D| = 500$). The results are provided in Table 4–Table 18, and they are the basis of the results reported in the main paper. The language codes are available in Table 1 (in the main paper). As mentioned in the main paper, all results are obtained with the two α-hyperparameters fixed to the following values: $\alpha_S = -0.25$, $\alpha_T = 0.15$, without any further fine-tuning. A more careful language pair-specific fine-tuning results in even higher performance for many language pairs.

In all tables, BASELINE refers to the best-performing weakly supervised projection-based approach *without* and *with* self-learning, as reported in a recent comparative study of Vulić et al. (2019); $5k$ and $1k$ denote the seed dictionary D size. The scores in bold indicate improvements over the BASELINE methods. All results are reported as MRR scores: the MRR score of $.xyz$ should be read as $xy.z\%$ (e.g., the score of .432 can be read as 43.2%).

(The actual tables with the full results in all BLI setups start on the next page.)

	Bulgarian: BG-													
	-CA	-EO	-ET	-EU	-FI	-HE	-HU	-ID	-KA	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.432	.327	.407	.250	.357	.361	.460	.283	.364	.205	.405	.398	.169	.349
BASELINE (self-learning, 5k)	.456	.370	.405	.296	.374	.368	.475	.325	.367	.215	.407	.446	.179	.374
POSTPROC (self-learning, 5k)	**.473**	**.419**	**.420**	**.302**	**.386**	**.392**	**.489**	**.330**	**.371**	.211	**.419**	**.462**	**.203**	**.379**
BASELINE (supervised, 1k)	.229	.147	.211	.070	.129	.112	.254	.116	.157	.054	.230	.163	.044	.133
BASELINE (self-learning, 1k)	.444	.357	.388	.279	.361	.345	.467	.314	.333	.186	.369	.441	.128	.357
POSTPROC (self-learning, 1k)	**.458**	**.408**	**.398**	**.286**	**.377**	**.376**	**.478**	**.321**	.329	**.188**	**.375**	**.458**	**.133**	**.362**

Table 4: All BLI scores (MRR) with Bulgarian (BG) as the source language.

	Catalan: CA-													
	-BG	-EO	-ET	-EU	-FI	-HE	-HU	-ID	-KA	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.396	.395	.356	.338	.329	.336	.431	.286	.309	.217	.366	.396	.196	.337
BASELINE (self-learning, 5k)	.414	.456	.352	.391	.356	.357	.449	.322	.302	.245	.343	.433	.218	.348
POSTPROC (self-learning, 5k)	**.434**	**.510**	**.359**	**.409**	**.359**	**.373**	**.454**	**.326**	**.322**	.242	.347	**.448**	**.234**	**.351**
BASELINE (supervised, 1k)	.212	.167	.165	.116	.110	.103	.210	.126	.101	.046	.144	.138	.035	.133
BASELINE (self-learning, 1k)	.395	.446	.300	.370	.319	.335	.435	.320	.253	.202	.295	.424	.142	.334
POSTPROC (self-learning, 1k)	**.413**	**.508**	**.309**	**.393**	**.321**	**.351**	**.439**	**.326**	**.274**	**.204**	**.306**	**.438**	**.146**	.332

Table 5: All BLI scores (MRR) with Catalan (CA) as the source language.

	Esperanto: EO-													
	-BG	-CA	-ET	-EU	-FI	-HE	-HU	-ID	-KA	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.367	.491	.334	.294	.329	.258	.400	.267	.281	.171	.343	.337	.107	.285
BASELINE (self-learning, 5k)	.410	.533	.342	.354	.363	.288	.426	.315	.296	.184	.384	.390	.117	.299
POSTPROC (self-learning, 5k)	**.428**	**.546**	**.353**	**.369**	**.372**	**.299**	**.432**	**.342**	**.311**	**.186**	**.404**	**.405**	**.124**	.292
BASELINE (supervised, 1k)	.152	.221	.136	.083	.080	.044	.145	.099	.078	.024	.120	.083	.017	.087
BASELINE (self-learning, 1k)	.385	.521	.314	.315	.328	.241	.411	.298	.255	.111	.358	.376	.056	.259
POSTPROC (self-learning, 1k)	**.404**	**.535**	**.318**	**.317**	.316	.235	.404	**.316**	**.271**	.092	**.368**	**.389**	**.061**	.251

Table 6: All BLI scores (MRR) with Esperanto (EO) as the source language.

	Estonian: ET-													
	-BG	-CA	-EO	-EU	-FI	-HE	-HU	-ID	-KA	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.393	.333	.271	.238	.430	.287	.432	.212	.258	.191	.360	.328	.168	.307
BASELINE (self-learning, 5k)	.404	.357	.307	.238	.443	.301	.459	.223	.251	.185	.358	.383	.178	.331
POSTPROC (self-learning, 5k)	**.433**	**.401**	**.352**	**.239**	**.447**	**.320**	**.471**	**.253**	.253	**.192**	**.380**	**.407**	**.205**	**.334**
BASELINE (supervised, 1k)	.200	.121	.116	.099	.200	.069	.188	.065	.095	.052	.179	.112	.041	.102
BASELINE (self-learning, 1k)	.381	.346	.297	.208	.437	.277	.449	.204	.215	.148	.337	.377	.108	.313
POSTPROC (self-learning, 1k)	**.415**	**.392**	**.337**	.200	**.446**	**.289**	**.461**	**.227**	**.224**	**.150**	**.356**	**.408**	.108	**.319**

Table 7: All BLI scores (MRR) with Estonian (ET) as the source language.

	Basque: EU-													
	-BG	-CA	-EO	-ET	-FI	-HE	-HU	-ID	-KA	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.292	.391	.245	.250	.233	.211	.259	.183	.197	.109	.242	.240	.095	.240
BASELINE (self-learning, 5k)	.310	.441	.277	.248	.270	.206	.283	.225	.189	.106	.237	.287	.094	.248
POSTPROC (self-learning, 5k)	**.332**	**.453**	**.324**	**.255**	**.276**	.207	**.302**	**.238**	.188	.108	.229	**.309**	**.119**	**.254**
BASELINE (supervised, 1k)	.120	.142	.077	.088	.048	.037	.077	.049	.059	.021	.071	.053	.018	.055
BASELINE (self-learning, 1k)	.276	.428	.253	.213	.247	.166	.266	.213	.147	.060	.169	.261	.056	.212
POSTPROC (self-learning, 1k)	**.294**	**.440**	**.292**	.209	.232	.144	.263	**.214**	.136	**.069**	.157	**.272**	**.059**	.201

Table 8: All BLI scores (MRR) with Basque (EU) as the source language.

	Finnish: FI-													
	-BG	-CA	-EO	-ET	-EU	-HE	-HU	-ID	-KA	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.379	.377	.284	.409	.220	.323	.456	.263	.275	.222	.390	.419	.171	.346
BASELINE (self-learning, 5k)	.397	.404	.320	.424	.271	.351	.474	.298	.289	.243	.405	.460	.168	.365
POSTPROC (self-learning, 5k)	**.423**	**.430**	**.386**	**.456**	**.302**	**.386**	**.477**	**.311**	**.329**	**.258**	**.434**	**.481**	**.196**	**.370**
BASELINE (supervised, 1k)	.174	.142	.077	.167	.054	.071	.226	.098	.084	.052	.158	.161	.028	.149
BASELINE (self-learning, 1k)	.381	.396	.304	.416	.235	.331	.463	.300	.270	.211	.389	.455	.107	.353
POSTPROC (self-learning, 1k)	**.409**	**.413**	**.372**	**.447**	**.259**	**.369**	**.466**	**.307**	**.303**	**.228**	**.424**	**.477**	**.112**	**.360**

Table 9: All BLI scores (MRR) with Finnish (FI) as the source language.

	Hebrew: HE-													
	-BG	-CA	-EO	-ET	-EU	-FI	-HU	-ID	-KA	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.397	.376	.248	.288	.225	.329	.375	.239	.213	.204	.309	.316	.173	.328
BASELINE (self-learning, 5k)	.378	.384	.278	.278	.211	.320	.393	.266	.217	.218	.301	.349	.192	.337
POSTPROC (self-learning, 5k)	**.401**	**.418**	**.307**	**.298**	.212	**.333**	**.402**	**.293**	.213	**.219**	.308	**.379**	**.238**	**.342**
BASELINE (supervised, 1k)	.180	.148	.087	.106	.065	.077	.135	.076	.067	.054	.105	.086	.042	.111
BASELINE (self-learning, 1k)	.360	.371	.252	.250	.182	.293	.383	.251	.188	.187	.254	.343	.114	.321
POSTPROC (self-learning, 1k)	**.381**	**.401**	**.280**	**.255**	.174	**.311**	**.388**	**.274**	.174	.184	**.255**	**.366**	**.131**	**.326**

Table 10: All BLI scores (MRR) with Hebrew (HE) as the source language.

	Hungarian: HU-													
	-BG	-CA	-EO	-ET	-EU	-FI	-HE	-ID	-KA	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.431	.443	.344	.423	.282	.397	.349	.338	.326	.259	.411	.406	.173	.372
BASELINE (self-learning, 5k)	.438	.477	.392	.433	.305	.407	.376	.374	.332	.285	.419	.441	.176	.380
POSTPROC (self-learning, 5k)	**.466**	**.495**	**.453**	**.457**	**.310**	**.418**	**.405**	**.403**	**.353**	**.293**	**.436**	**.457**	**.194**	**.387**
BASELINE (supervised, 1k)	.241	.221	.125	.196	.094	.168	.098	.147	.112	.063	.183	.149	.026	.184
BASELINE (self-learning, 1k)	.427	.467	.369	.413	.274	.400	.356	.377	.306	.268	.381	.423	.113	.374
POSTPROC (self-learning, 1k)	**.458**	**.484**	**.431**	**.443**	**.276**	**.410**	**.385**	**.406**	**.331**	**.270**	**.401**	**.447**	**.126**	**.377**

Table 11: All BLI scores (MRR) with Hungarian (HU) as the source language.

	Indonesian: ID-													
	-BG	-CA	-EO	-ET	-EU	-FI	-HE	-HU	-KA	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.281	.300	.247	.281	.173	.233	.290	.349	.222	.193	.260	.294	.218	.316
BASELINE (self-learning, 5k)	.287	.323	.274	.266	.220	.269	.295	.345	.200	.197	.242	.320	.241	.326
POSTPROC (self-learning, 5k)	**.307**	**.333**	**.303**	.273	**.225**	**.270**	**.298**	**.360**	.205	**.203**	.242	**.335**	**.256**	**.328**
BASELINE (supervised, 1k)	.121	.114	.092	.115	.038	.053	.093	.129	.063	.062	.086	.081	.052	.152
BASELINE (self-learning, 1k)	.258	.316	.254	.213	.187	.250	.264	.337	.140	.175	.152	.309	.226	.319
POSTPROC (self-learning, 1k)	**.280**	**.327**	**.282**	**.221**	**.197**	**.252**	**.271**	**.346**	.131	**.184**	.149	**.325**	.225	**.322**

Table 12: All BLI scores (MRR) with Indonesian (ID) as the source language.

	Georgian: KA-													
	-BG	-CA	-EO	-ET	-EU	-FI	-HE	-HU	-ID	-KO	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.372	.297	.243	.282	.217	.292	.245	.308	.169	.154	.327	.214	.127	.257
BASELINE (self-learning, 5k)	.376	.320	.265	.293	.216	.318	.251	.326	.172	.143	.340	.253	.139	.275
POSTPROC (self-learning, 5k)	**.412**	**.355**	**.307**	**.300**	**.218**	**.331**	**.270**	**.343**	**.200**	.154	**.342**	**.281**	**.153**	**.280**
BASELINE (supervised, 1k)	.153	.088	.083	.112	.068	.065	.046	.103	.048	.036	.138	.048	.025	.091
BASELINE (self-learning, 1k)	.352	.305	.248	.271	.172	.306	.213	.308	.155	.103	.317	.238	.077	.255
POSTPROC (self-learning, 1k)	**.378**	**.341**	**.283**	**.279**	**.174**	**.308**	**.233**	**.323**	**.177**	.098	**.321**	**.260**	**.078**	.249

Table 13: All BLI scores (MRR) with Georgian (KA) as the source language.

	Korean: KO-													
	-BG	-CA	-EO	-ET	-EU	-FI	-HE	-HU	-ID	-KA	-LT	-NO	-TH	-TR
BASELINE (supervised, 5k)	.190	.183	.083	.145	.102	.206	.166	.238	.142	.112	.156	.150	.076	.213
BASELINE (self-learning, 5k)	.289	.283	.176	.242	.170	.273	.257	.326	.210	.178	.241	.256	.174	.278
POSTPROC (self-learning, 5k)	**.324**	**.330**	**.217**	**.247**	.153	**.310**	**.281**	**.367**	**.264**	**.180**	.239	**.313**	**.199**	**.301**
BASELINE (supervised, 1k)	.093	.078	.045	.059	.045	.066	.048	.096	.060	.039	.053	.047	.038	.085
BASELINE (self-learning, 1k)	.245	.253	.110	.191	.108	.266	.232	.343	.206	.122	.150	.244	.089	.279
POSTPROC (self-learning, 1k)	**.268**	**.274**	**.134**	**.193**	.106	**.271**	**.239**	**.348**	**.236**	**.117**	**.152**	**.264**	**.102**	**.284**

Table 14: All BLI scores (MRR) with Korean (KO) as the source language.

	Lithuanian: LT-													
	-BG	-CA	-EO	-ET	-EU	-FI	-HE	-HU	-ID	-KA	-KO	-NO	-TH	-TR
BASELINE (supervised, 5k)	.462	.353	.317	.394	.236	.368	.299	.395	.184	.284	.168	.304	.162	.296
BASELINE (self-learning, 5k)	.437	.363	.348	.383	.222	.385	.316	.413	.191	.304	.160	.336	.168	.319
POSTPROC (self-learning, 5k)	**.470**	**.408**	**.406**	**.400**	.233	**.394**	**.338**	**.426**	**.220**	.300	.160	**.372**	**.205**	**.326**
BASELINE (supervised, 1k)	.256	.138	.102	.190	.085	.143	.073	.159	.058	.097	.040	.081	.030	.097
BASELINE (self-learning, 1k)	.408	.345	.332	.361	.181	.380	.286	.399	.168	.288	.109	.322	.094	.302
POSTPROC (self-learning, 1k)	**.438**	**.387**	**.388**	**.382**	**.191**	**.390**	**.306**	**.412**	**.195**	.282	**.117**	**.355**	**.109**	**.305**

Table 15: All BLI scores (MRR) with Lithuanian (LT) as the source language.

	Norwegian: NO-													
	-BG	-CA	-EO	-ET	-EU	-FI	-HE	-HU	-ID	-KA	-KO	-LT	-TH	-TR
BASELINE (supervised, 5k)	.394	.424	.323	.389	.261	.396	.319	.441	.306	.291	.220	.366	.188	.325
BASELINE (self-learning, 5k)	.422	.457	.377	.395	.328	.419	.353	.452	.340	.298	.250	.351	.197	.341
POSTPROC (self-learning, 5k)	**.441**	**.474**	**.425**	**.411**	**.345**	**.424**	**.381**	**.455**	**.354**	**.315**	**.257**	**.367**	**.227**	**.346**
BASELINE (supervised, 1k)	.203	.198	.128	.172	.075	.153	.078	.206	.132	.088	.057	.132	.032	.123
BASELINE (self-learning, 1k)	.411	.444	.374	.371	.300	.412	.336	.443	.339	.268	.228	.315	.140	.332
POSTPROC (self-learning, 1k)	**.433**	**.466**	**.419**	**.389**	**.313**	**.417**	**.366**	**.445**	**.352**	**.279**	**.236**	**.332**	.136	**.336**

Table 16: All BLI scores (MRR) with Norwegian (NO) as the source language.

	Thai: TH-													
	-BG	-CA	-EO	-ET	-EU	-FI	-HE	-HU	-ID	-KA	-KO	-LT	-NO	-TR
BASELINE (supervised, 5k)	.210	.134	.087	.186	.094	.173	.173	.178	.141	.116	.112	.214	.162	.177
BASELINE (self-learning, 5k)	.174	.123	.073	.164	.093	.167	.203	.160	.170	.126	.097	.215	.147	.160
POSTPROC (self-learning, 5k)	.176	**.145**	.068	.168	**.098**	**.178**	.176	**.188**	**.203**	**.136**	**.118**	**.218**	.143	.170
BASELINE (supervised, 1k)	.049	.027	.021	.070	.029	.032	.057	.044	.044	.034	.040	.084	.029	.052
BASELINE (self-learning, 1k)	.108	.084	.036	.128	.057	.094	.152	.111	.168	.073	.065	.145	.098	.121
POSTPROC (self-learning, 1k)	**.112**	**.104**	**.049**	.120	.049	**.104**	.150	**.127**	**.192**	**.079**	**.078**	**.151**	**.107**	**.125**

Table 17: All BLI scores (MRR) with Thai (TH) as the source language.

	Turkish: TR-													
	-BG	-CA	-EO	-ET	-EU	-FI	-HE	-HU	-ID	-KA	-KO	-LT	-NO	-TH
BASELINE (supervised, 5k)	.344	.360	.215	.307	.230	.294	.319	.378	.336	.205	.196	.295	.311	.170
BASELINE (self-learning, 5k)	.351	.376	.238	.309	.244	.322	.323	.397	.370	.229	.214	.280	.346	.183
POSTPROC (self-learning, 5k)	**.378**	**.405**	**.291**	**.328**	**.252**	**.338**	**.361**	**.413**	**.395**	**.261**	**.226**	**.298**	**.369**	**.210**
BASELINE (supervised, 1k)	.150	.133	.052	.112	.062	.093	.076	.167	.131	.053	.050	.099	.073	.028
BASELINE (self-learning, 1k)	.327	.364	.204	.274	.209	.310	.301	.398	.363	.201	.194	.215	.344	.137
POSTPROC (self-learning, 1k)	**.361**	**.394**	**.259**	**.289**	**.217**	**.326**	**.336**	**.411**	**.390**	**.245**	**.200**	**.234**	**.368**	**.142**

Table 18: All BLI scores (MRR) with Turkish (TR) as the source language.