Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces

Goran Glavaš[1] **and Ivan Vulić**[2]
[1] Data and Web Science Group, University of Mannheim, Germany
[2]Language Technology Lab, TAL, University of Cambridge, UK
goran@informatik.uni-mannheim.de
iv250@cam.ac.uk

Abstract

We present INSTAMAP, an instance-based method for learning projection-based cross-lingual word embeddings. Unlike prior work, it deviates from learning a single global linear projection. INSTAMAP is a non-parametric model that learns a non-linear projection by iteratively: (1) finding a globally optimal rotation of the source embedding space relying on the Kabsch algorithm, and then (2) moving each point along an instance-specific translation vector estimated from the translation vectors of the point's nearest neighbours in the training dictionary. We report performance gains with INSTAMAP over four representative state-of-the-art projection-based models on bilingual lexicon induction across a set of 28 diverse language pairs. We note prominent improvements, especially for more distant language pairs (i.e., languages with non-isomorphic monolingual spaces).

1 Introduction and Motivation

Induction of cross-lingual word embeddings (CLWEs) (Vulić et al., 2011; Mikolov et al., 2013; Xing et al., 2015; Smith et al., 2017; Artetxe et al., 2018) has been one of the key mechanisms for enabling multilingual modeling of meaning and facilitating cross-lingual transfer for downstream NLP tasks. Even though CLWEs are recently being contested in cross-lingual downstream transfer by pretrained multilingual language models (Pires et al., 2019; Conneau et al., 2020; Artetxe et al., 2019; Wu and Dredze, 2019; Wu et al., 2020), they are still paramount in word-level translation, that is, bilingual lexicon induction (BLI).

While earlier work focused on joint induction of multilingual embeddings from multilingual corpora, relying on word- (Klementiev et al., 2012; Kočiskỳ et al., 2014; Gouws and Søgaard, 2015), sentence- (Zou et al., 2013; Hermann and Blunsom, 2014; Luong et al., 2015; Coulmance et al., 2015; Levy et al., 2017), or document-level (Søgaard et al., 2015; Mogadala and Rettinger, 2016; Vulić and Moens, 2016) alignments, most recent efforts focus on post-hoc alignment of independently trained monolingual embeddings (the so-called *projection* or *mapping* approaches) (Smith et al., 2017; Artetxe et al., 2018; Conneau et al., 2018; Joulin et al., 2018; Patra et al., 2019, *inter alia*).

Despite some recent evidence that joint CLWE induction may lead to better bilingual spaces (Ormazabal et al., 2019), projection-based methods still dominate the field (Hoshen and Wolf, 2018; Ruder et al., 2018; Nakashole, 2018; Grave et al., 2019; Zhang et al., 2019, *inter alia*) due to their conceptual attractiveness: they operate on top of vectors produced with any embedding model and need at most a few thousand word pairs of supervision (Glavaš et al., 2019; Vulić et al., 2019).

Most projection-based CLWE models induce bilingual spaces by orthogonally projecting one monolingual space to another. Since orthogonal projections do not affect the topology of the source space, the performance of these methods is bound by the degree of isomorphism of the two monolingual spaces. Yet, evidence suggests that monolingual spaces, especially those of etymologically and typologically distant languages, are far from isomorphic (Søgaard et al., 2018; Vulić et al., 2019; Patra et al., 2019). What is more, unsupervised CLWE models (Conneau et al., 2018; Artetxe et al., 2018; Alvarez-Melis and Jaakkola, 2018; Hoshen and Wolf, 2018, *inter alia*), which additionally exploit the isomorphism assumption when inducing initial translation dictionaries, have been shown to yield near-zero BLI results for pairs of distant languages (Søgaard et al., 2018; Vulić et al., 2019). Following these theoretical limitations of effectiveness of orthogonal mapping between non-isomorphic spaces, Joulin et al. (2018) and Patra

et al. (2019) relax the orthogonality constraint and report BLI improvements. These models, however, still learn only a linear transformation, i.e., an oblique projection matrix. While oblique projections may scale or skew the source space, there still exists a strong topological similarity between the original space and its oblique projection.

In this work, we deviate from learning a linear projection matrix (i.e., a parametric model) and propose a non-parametric model which translates vectors by estimating instance-specific geometric translations. Our method, INSTAMAP, iteratively (1) applies the Kabsch algorithm (Horn, 1987) on the full training dictionary to learn a globally optimal rotation of the source space w.r.t. the target space; and then (2) translates each point along the instance-specific translation vector, which we compute from the translation vectors of the point's nearest neighbours from the training dictionary.

We extensively evaluate INSTAMAP on the benchmark BLI dataset (Glavaš et al., 2019) encompassing 28 diverse language pairs. Our results show the non-linear mappings with INSTAMAP to be substantially more robust than linear projections, both orthogonal (Smith et al., 2017; Artetxe et al., 2018) and oblique (Joulin et al., 2018; Patra et al., 2019). We also show that, unlike INSTAMAP, oblique projection models – RCSLS (Joulin et al., 2018) and BLISS (Patra et al., 2019) – cannot surpass the performance of the best-performing orthogonal projection model VecMap (Artetxe et al., 2018) for distant languages (i.e., for low isomorphicity). Finally, we report additional significant gains by applying INSTAMAP on top of VecMap.

2 Instance-Based Mapping

The core idea of INSTAMAP is illustrated in Figure 1. We iteratively: (1) use the entire training dictionary to learn a single global rotation matrix and then (2) perform an instance-based computation of translation vectors.

2.1 Globally Optimal Rotation

Let $\mathbf{X}$ and $\mathbf{Y}$ be monolingual embedding spaces of the source and target language, respectively, and let $D = \{(w^i_{L1}, w^i_{L2})\}, i = 1 \dots N$, be the training dictionary. We first transform each of the two spaces by (independently) performing a full PCA transformation (i.e., no dimensionality reduction): this way we represent vectors in each of the spaces as combinations of linearly uncorrelated principal components of that space, which facilitates the learning of the optimal rotation between the spaces. Let $\mathbf{X}_D = \{\mathbf{x}^i_{L1}\}^N_{i=1} \subset \mathbf{X}$ and $\mathbf{Y}_D = \{\mathbf{y}^i_{L2}\}^N_{i=1} \subset \mathbf{Y}$ be the dictionary-aligned subsets of the two monolingual spaces. We aim to learn the optimal rotation matrix between $\mathbf{X}$ and $\mathbf{Y}$, i.e., the matrix $\mathbf{W}_R$ that minimizes the sum of square distances between the source vector projections and corresponding target vectors, $W_R = \arg\min_W \|\mathbf{X}_D\mathbf{W} - \mathbf{Y}_D\|$. If we constrain $\mathbf{W_R}$ to be orthogonal, the optimal solution is obtained by solving the Procrustes problem (Schönemann, 1966) – adopted by most projection-based CLWE models (Smith et al., 2017; Conneau et al., 2018; Artetxe et al., 2018). However, our aim is to avoid introducing the orthogonal constraint and learn only the optimal rotation between the spaces. To this end, we use the Kabsch algorithm (Horn, 1987), which computes the optimal rotation matrix $\mathbf{W}_R$ as follows:

$$\mathbf{W}_R = \mathbf{V}\,\mathbf{I}_R\,\mathbf{U}^T, \text{with} \quad (1)$$

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = SVD(\mathbf{X}_D^T\mathbf{Y}_D), \quad (2)$$

where $\mathbf{I}_R$ is a modification of the identity matrix, in which the last element (i.e., last row, last column) is not 1, but rather the determinant of $\mathbf{V}\mathbf{U}^T$. Upon obtaining $\mathbf{W}_R$, we rotate $\mathbf{X}$ w.r.t. $\mathbf{Y}$, $\mathbf{X}_R = \mathbf{X}\mathbf{W}_R$.

2.2 Instance-Specific Translations

We then perform localized, instance-specific translations in a rotationally-aligned bilingual space. For each point from both $\mathbf{X}_R$ and $\mathbf{Y}$, we compute a "personalized" translation vector, as the weighted average of the translation vectors of its closest dictionary entries. That is, for some vector $\mathbf{x} \in \mathbf{X}_R$ let $\mathbf{x}_1, \dots, \mathbf{x}_K$ be the set of K vectors from $\mathbf{X}_D\mathbf{W}_R$ (corresponding to words w^1_{L1}, $w^2_{L1}, \dots, w^K_{L1}$ in D) which are closest to $\mathbf{x}$ in terms of cosine similarity and let $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_K$ be the vectors of the corresponding dictionary translations w^1_{L2}, $w^2_{L2}, \dots, w^K_{L2}$ from D from the target language space. We then compute the instance-based translation of $\mathbf{x}$, $\mathbf{x}'$, as follows:

$$\mathbf{x}' = \mathbf{x} + \frac{\sum_{k=1}^K \cos(\mathbf{x}, \mathbf{x}_k) \cdot (\mathbf{y}_k - \mathbf{x}_k)}{\sum_{k=1}^K \cos(\mathbf{x}, \mathbf{x}_k)} \quad (3)$$

We perform an instance-specific translation of the vectors from $\mathbf{Y}$ analogously. Let $\mathbf{y}_1, \dots, \mathbf{y}_K$ be the set of vectors from $\mathbf{Y}_D$ that are closest to some vector $\mathbf{y} \in \mathbf{Y}$. The translation $\mathbf{y}'$ is then as follows:

$$\mathbf{y}' = \mathbf{y} - \frac{\sum_{k=1}^K \cos(\mathbf{y}, \mathbf{y}_k) \cdot (\mathbf{y}_k - \mathbf{x}_k)}{\sum_{k=1}^K \cos(\mathbf{y}, \mathbf{y}_k)} \quad (4)$$

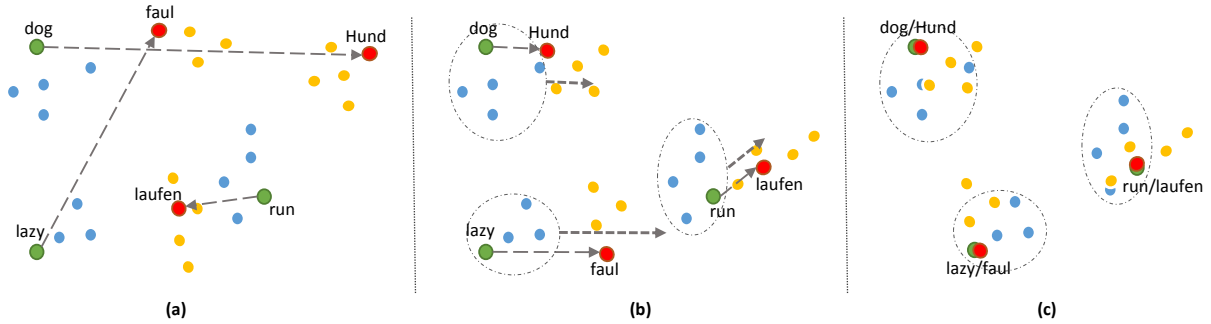


Figure 1: Illustration of INSTAMAP: **(a)** unaligned monolingual embedding spaces (EN – blue; DE – yellow) with dictionary alignments D (EN – green; DE – red); **(b)** rotation-aligned spaces: rotation matrix is learned on the whole dictionary D; **(c)** INSTAMAP bilingual space: each point's translation vector (depicted in (b)) is computed from translation vectors of nearest entries in D; *Nota bene*: for simplicity, each point in illustration (figure (b)) inherits the translation vector of the nearest dictionary entry; in the actual algorithm, however, the translation vector is computed as weighted average of translation vectors of K nearest neighbours in D (see the description in §2.2).

Because we compute a different translation vector for each point in both vector spaces, the final mapping function between the two spaces is globally non-linear. Also, being based on K nearest neighbours in the training dictionary D, INSTAMAP is, unlike all other projection-based CLWE models, a non-parametric model (i.e., the number of model parameters is not fixed, it depends on the number of entries in the training dictionary D).

2.3 Training Dictionary Expansion

We repeat the two steps – global rotation and instance-based translation – aiming to obtain an iterative refinement of the non-linear mapping between the two spaces. Following the established practice found in other iterative models (Conneau et al., 2018; Artetxe et al., 2018), we augment the training dictionary for the next iteration with the mutual nearest neighbours in the bilingual space induced in the previous iteration. Intuitively, with INSTAMAP being a non-parametric model, we expect it to benefit more from the dictionary augmentation than the parametric projection models, which have been shown to saturate in performance when training dictionaries exceed 5K-10K translation pairs (Vulić and Korhonen, 2016; Glavaš et al., 2019).

3 Evaluation

We evaluate INSTAMAP on bilingual lexicon induction, the standard task for evaluating CLWEs.

3.1 Experimental Setup

Data. We evaluate on the BLI benchmark dataset introduced by Glavaš et al. (2019), containing 28 pairs between eight diverse languages: English (EN), German (DE), Italian (IT), French (FR), Russian (RU), Croatian (HR), Turkish (TR), and Finnish (FI).[1] Comprising both close and distant language pairs, this dataset allows us to compare model performance in settings with varying degree of isomorphism between monolingual spaces. We start from monolingual FastText vectors trained on Wikipedias of respective languages,[2] with vocabularies trimmed to the 200K most frequent words.

Baselines. We compare INSTAMAP to the baseline orthogonal projection – solution to the Procrustes problem (PROC), and three state-of-the-art projection-based models: (1) VecMap (Artetxe et al., 2018) emerged in recent comparative evaluations (Glavaš et al., 2019; Vulić et al., 2019) as the best-performing orthogonal-projection model; (2) RCSLS (Joulin et al., 2018) learns an oblique (i.e., non-orthogonal) projection and yields best performance overall in a recent comparative evaluation (Glavaš et al., 2019); (3) BLISS (Patra et al., 2019) combines an orthogonal projection objective with an objective based on adversarial learning, inducing a weakly-orthogonal projection matrix.

Model Variants and Hyperparameter Tuning. We evaluate two variants of INSTAMAP: (1) the base model is applied directly on unaligned monolingual vector spaces; (2) IM ∘ VM is the variant in which we apply INSTAMAP on top of the bilingual space induced with VecMap (Artetxe et al., 2018): because VecMap induces an orthogonal projection, the topologies of the monolingual subspaces of the

[1] We use the training dictionaries with 5K instances.
[2] https://fasttext.cc/

Model	Projection	ALL	EN-*	No-EN	EASY	HARD	Best LPs
PROC	Orthogonal	32.26	38.07	29.97	50.96	20.93	DE-RU, DE-IT, DE-FR, IT-FR, EN-DE
VECMAP	Orthogonal	36.08	42.09	33.77	53.69	24.24	HR-IT, DE-IT, FI-HR, FI-FR, HR-FR
RCSLS	Oblique	35.31	41.94	32.75	53.31	23.78	DE-RU, DE-FI, DE-HR, DE-TR, DE-FR
BLISS	Oblique	33.78	44.62	30.04	49.92	21.07	EN-RU, EN-TR, HR-RU, EN-HR, EN-FR
INSTAMAP	Non-linear	36.94	42.42	34.87	53.99	25.71	DE-HR, DE-TR, DE-RU, FI-IT, DE-FR
IM ∘ VM	Non-linear	**38.69**	**44.82**	**36.43**	**55.01**	**27.72**	DE-HR, TR-HR, TR-FI, DE-FI, DE-TR

Table 1: BLI results aggregated over diverse language pairs. Setups: (a) ALL – all 28 language pairs from (Glavaš et al., 2019) (b) EN-* – 7 language pairs with English as the source language; (c) EASY – 6 (20%) least difficult language pairs (EN-DE, EN-IT, EN-FR, IT-FR, DE-IT, DE-FR), according to average ranking of all models in evaluation; (d) HARD – 6 (20%) most difficult language pairs (TR-HR, DE-TR, TR-FI, TR-RU, FI-HR, DE-HR). (e) BEST LPs – 5 language pairs for which each model yields best relative performance compared to other models.

VecMap bilingual space are preserved compared to respective original monolingual spaces – this holds promise of no undesirable side-effects originating from the composition. INSTAMAP has only two hyperparameters:[3] the number of nearest neighbours K from D, and the number of algorithm iterations T. We identified, via fixed-split cross-validation on the training dictionaries, that configuration $K = 70$ and $T = 4$ works best for most language pairs.[4]

3.2 Results

We show BLI performance ($P@1$), aggregated over several different sets of language pairs, in Table 1.[5] Overall, INSTAMAP significantly outperforms all competing models[6] Somewhat surprisingly, VecMap, which induces an orthogonal projection (i.e., more strongly relies on the assumption of isomorphism), significantly outperforms RCSLS and BLISS, models that relax the orthogonality constraint and induce oblique linear projections. Only INSTAMAP, by removing the constraint of having a global linear projection altogether and by inducing a non-linear mapping, is able to consistently yield improvements over the orthogonal projection (VecMap). What is more, the IM ∘ VM composition yields even larger performance gains.

Analysis of results across different groups of language pairs identifies INSTAMAP as particularly beneficial for pairs of distant languages (setups No-EN and HARD) and languages with least reliable monolingual vectors (TR, HR). For example, while INSTAMAP alone and IM ∘ VM yield gains of 0.9 and 2.6 points, respectively, w.r.t. VecMap across ALL language pairs, these gaps widen to 1.5 and 3.5 points on most challenging language pairs (HARD). In contrast, BLISS, a model specifically tailored to improve the mappings between non-isomorphic spaces, appears to be robust only on pairs of close languages (e.g., HR-RU) and pairs involving EN (setup EN-*). It exhibits barely any improvement over the baseline orthogonal projection (PROC) on distant language pairs (HARD) and a significant degradation w.r.t. VecMap, a state-of-the-art model based on orthogonal projection. RCSLS is more robust than BLISS on difficult language pairs, but still performs worse than VecMap.

Further Analysis. We further analyze the performance of INSTAMAP (applied on top of VecMap) with respect to: (1) size of the training dictionary $|D|$ and (2) number of nearest dictionary neighbours K. We analyze the performance of IM∘VM for three language pairs with lowest BLI scores: DE-TR, TR-FI, and TR-HR. We prepare dictionaries with 2.5K to 12.5K entries (with a 2.5K step), following steps described in (Glavaš et al., 2019).[7] Figure 2 shows the performance for different training dictionary sizes. We can see that adding INSTAMAP on top of VecMap yields stable improvements for all dictionary sizes. On the one hand, this shows that INSTAMAP is equally helpful for any number of available word translations. On the other hand, since InstaMap is not constrained to learning a single global projection, we hoped to see bigger gains for larger dictionaries, but this

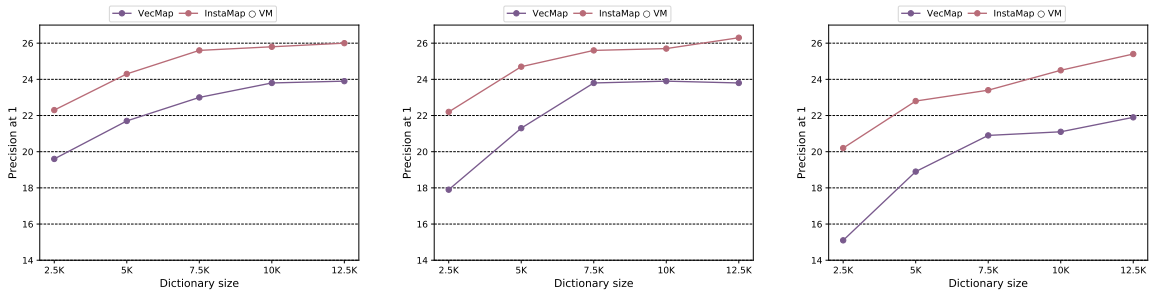
[3] Competing models – VecMap, RCSLS, and BLISS – all come with much larger sets of hyperparameters.

[4] For some pairs other configurations yield slightly better results: for simplicity, we report the results with base configuration $K = 70; T = 4$ for all pairs.

[5] We provide detailed results for each of 28 language language pairs in the supplemental material.

[6] Non-parametric shuffling test (Yeh, 2000) with the Bonferroni correction: $\alpha < 0.05$ in comparison with VecMap and $\alpha < 0.01$ in comparison with other models.

[7] We translate 20K most frequent EN words to DE, TR, FI, and HR and keep for each language pair only word pairs (1) found in respective monolingual FastText vocabularies, (2) not present in the 2K test dictionaries from (Glavaš et al., 2019).



(a) German-Turkish (b) Turkish-Finnish (c) Turkish-Croatian

Figure 2: Comparison of performance between VECMAP and INSTAMAP applied on top it (IM ○ VM) for different sizes of the training dictionary (from 2.5K word pairs to 12.5K word pairs in steps of 2.5K pairs).

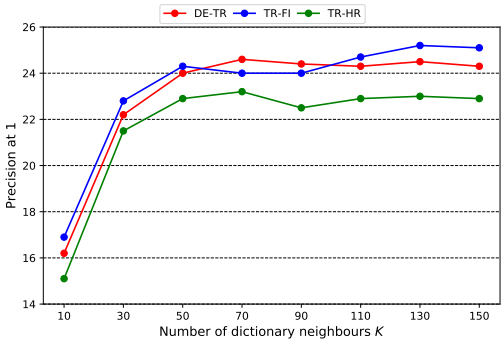


Figure 3: *InstaMap* (IM ○ VM) performance w.r.t the number of nearest dictionary neighbours K. Results shown for three language pairs – DE-TR, TR-FI, and TR-HR – and $T = 3$ INSTAMAP iterations.

is not the case. With larger dictionaries, we are more likely to find more semantically similar dictionary neighbours for each word – and this should lead to better performance. We speculate, however, that larger dictionaries also increase the likelihood of selecting spurious neighbours due to hubness (Dinu et al., 2015; Conneau et al., 2018) and that this cancels out the positive effect promised by having more candidates to choose the neighbours from. This could perhaps be remedied by using hubness-aware similarity scores like CSLS (Conneau et al., 2018) instead of simple cosine similarity.

Figure 3 illustrates how INSTAMAP performance (on top of VecMap, i.e., IM ○ VM) varies with different values for the number of dictionary neighbours K. The best performance is typically reached for values of K between 50 and 90 and there are no further improvements for larger values of K (TR-FI, where $K = 130$ gives the best score, is an exception). For very small K performance drops are substantial and here INSTAMAP even degrades the quality of the input space produced by VecMap. We believe this happens because INSTAMAP in this case has too few dictionary neighbours to accurately model the meaning of any given word and, in turn, compute a reliable mapping vector.

4 Conclusion

We have proposed INSTAMAP, a simple and effective approach for improving the post-hoc cross-lingual alignment between non-isomorphic monolingual embedding spaces. Unlike existing projection-based CLWE induction models, which learn a global linear projection matrix, INSTAMAP couples global rotation with instance-specific translations. This way, we learn a globally non-linear projection. Our experiments show that (1) INSTAMAP significantly outperforms four state-of-the-art projection-based CLWE models on a benchmark BLI dataset with 28 language pairs and (2) that it yields largest improvements for pairs of distant languages with a lower degree of isomorphism between their respective monolingual spaces. We plan to extend this work in two directions. First, we will explore mechanisms for instance-specific translation that are more sophisticated than the aggregation of translation vectors of nearest dictionary neighbours. Second, we plan to couple instance-based mapping with other informative features (e.g., character-level features) in classification-based BLI frameworks (Heyman et al., 2017; Karan et al., 2020). The INSTAMAP code is available at: https://github.com/codogogo/instamap.

Acknowledgments

We thank the anonymous reviewers for their insightful suggestions. GG is supported by the Eliteprogramm of the Baden-Württemberg Stiftung (AGREE grant). IV is supported by the ERC Consolidator Grant LEXICAL (no 648909).

References

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

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

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

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

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of EMNLP*, pages 1109–1113.

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

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

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

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *Proceedings of AISTATS*, pages 1880–1890.

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

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 the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1085–1095.

Berthold K.P. Horn. 1987. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642.

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.

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.

Tomáš Kočiskỳ, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of ACL*, pages 224–229.

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.

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.

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

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of NAACL-HLT*, pages 692–702.

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

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

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of ACL*, pages 184–193.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*, pages 4996–5001.

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.

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.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of ACL*, pages 479–484.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP*, pages 4398–4409.

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

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of ACL*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP*, pages 833–844.

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.

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

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.

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.

Model	Proj.	EN-DE	EN-TR	EN-FI	EN-HR	EN-RU	EN-IT	EN-FR	DE-TR	DE-FI	DE-HR	DE-RU	DE-IT	DE-FR	TR-FI
PROC	Orth.	45.6	24.1	30.8	24.3	36.2	54.8	58.2	20.8	27.1	24.2	32.9	43.2	42.5	19.7
VECMAP	Orth.	48.5	30.4	35.4	29.8	38.5	57.3	61.2	23.2	28.8	27.1	31.6	46.3	45.3	23.1
RCSLS	Obliq.	49.0	28.8	34.8	28.1	41.9	57.6	60.6	24.1	30.2	27.3	36.8	45.0	45.2	22.8
BLISS	Obliq.	47.0	32.7	35.8	30.6	**45.9**	58.9	**63.9**	17.5	20.3	17.4	20.5	30.9	33.3	21.6
INSTAMAP	Non-lin.	47.6	30.6	33.6	29.6	40.6	59.1	61.2	25.7	31.0	29.1	36.4	46.8	**47.4**	24.9
IM ∘ VM	Non-lin.	**49.3**	**33.0**	**37.7**	**32.5**	42.9	**59.8**	63.1	**26.8**	**33.5**	**31.1**	**37.6**	**47.4**	46.9	**27.4**

Table 2: BLI results detailed over the first batch of 14 language pairs.

Model	Proj.	TR-HR	TR-RU	TR-IT	TR-FR	FI-HR	FI-RU	FI-IT	FI-FR	HR-RU	HR-IT	HR-FR	RU-IT	RU-FR	IT-FR
PROC	Orth.	18.4	20.9	25.2	25.9	21.4	25.4	27.3	28.1	28.8	27.9	29.2	40.1	38.6	61.5
VECMAP	Orth.	21.9	23.5	30.6	31.6	26.7	30.3	32.6	34.0	33.1	35.4	34.9	42.8	42.9	63.5
RCSLS	Obliq.	20.2	24.3	28.2	29.5	23.8	29.1	29.9	30.8	31.9	30.7	32.5	41.8	41.0	62.6
BLISS	Obliq.	21.6	23.1	28.2	29.0	25.3	29.5	30.4	31.1	34.8	32.3	32.5	42.9	43.3	**65.6**
INSTAMAP	Non-lin.	23.9	24.4	31.4	31.7	26.2	30.6	35.0	34.4	33.4	35.1	34.2	44.4	44.5	61.9
IM ∘ VM	Non-lin.	**26.1**	**26.6**	**32.0**	**34.5**	**28.4**	**31.8**	**35.4**	**35.8**	**36.3**	**36.4**	**36.6**	**44.9**	**46.0**	63.5

Table 3: BLI results detailed over the second batch of 14 language pairs.