24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Refinement of Unsupervised Cross-Lingual Word Embeddings

Magdalena Biesialska [1] and **Marta R. Costa-jussà** [1]

Abstract. Cross-lingual word embeddings aim to bridge the gap between high-resource and low-resource languages by allowing to learn multilingual word representations even without using any direct bilingual signal. The lion's share of the methods are projection-based approaches that map pre-trained embeddings into a shared latent space. These methods are mostly based on the orthogonal transformation, which assumes language vector spaces to be isomorphic. However, this criterion does not necessarily hold, especially for morphologically-rich languages. In this paper, we propose a self-supervised method to refine the alignment of unsupervised bilingual word embeddings. The proposed model moves vectors of words and their corresponding translations closer to each other as well as enforces length- and center-invariance, thus allowing to better align cross-lingual embeddings. The experimental results demonstrate the effectiveness of our approach, as in most cases it outperforms state-of-the-art methods in a bilingual lexicon induction task.

1 INTRODUCTION

There are roughly 7000 languages around the world [17], and thus multilingual Natural Language Processing (NLP) has been a long-standing goal. Yet, NLP systems nowadays mainly support only the English language. This stems from a limited number of available parallel corpora even for resource-rich languages. The necessity to rely on bilingual data poses a great constraint on the development of multilingual NLP systems.

Cross-lingual word embeddings (CLEs) aim to bridge the gap between high-resource and low-resource languages by enabling to learn multi-lingual word representations even without any parallel data. More specifically, CLEs are representations of words in different languages, trained independently on monolingual corpora, and subsequently mapped into a shared vector space via linear transformation. Not surprisingly, CLEs have been attracting a lot of attention lately, as they allow to compare a word's meaning between languages and enable cross-lingual transfer learning [22]. These properties are beneficial for resource-rich languages, but are even more desirable in low-resource scenarios. Hence, this makes CLEs useful in downstream NLP tasks, such as: bilingual lexicon induction, neural machine translation, document classification, and information retrieval, among others.

In fact, there exist various methods to obtain CLEs, where a key differentiator is the nature and amount of a bilingual signal provided during training.

Supervised. Early methods [19, 11] leveraged large, prepared in advance, bilingual dictionaries to learn cross-lingual embedding mappings. Later it was shown that the number of seed word translations can be reduced considerably [26], however, the requirement of bilingual supervision has remained the same.

Weakly supervised. These bootstrapping approaches rely on typically small seed lexicons. In particular, CLE models that exploit a weak supervision, use initial bilingual seeds based on: cognates [24], identical words [25] or shared numerals [4].

Unsupervised. The most recent line of research [28, 6, 8], allows to learn CLEs without the need of any bilingual signal. Interestingly, CLEs trained solely on monolingual corpora are reported to demonstrate performance on a par with or even outperform supervised methods [6, 8, 1]. Importantly, Grave *et al.* [14] observe that refinement methods significantly improve the quality of weakly supervised and unsupervised CLE models.

In this work, we make a number of contributions. Firstly, we introduce a method for a self-supervised refinement of unsupervised CLEs. In contrast with existing approaches, our method is fully unsupervised and leverages a small self-learned seed lexicon. In addition, to the best of our knowledge, we are the first to apply a self-supervised refinement method to the state-of-the-art[2] unsupervised CLE model proposed by Artetxe *et al.* [6]. Secondly, in this work, we address the problem of imperfect isomorphism in embedding vector spaces. Lastly, through the evaluation of our approach on a standard bilingual dictionary induction benchmark, we show that our method improves the word translation accuracy for almost all investigated language pairs.

2 RELATED WORK

2.1 Isomorphic Vector Spaces and Orthogonal Transformations

According to the popular claim stated by Mikolov *et al.* [19], an alignment between word vector spaces representing two different languages is possible, because same concepts in different languages bear similar statistical properties, and thus vector spaces of these languages can be considered isomorphic. Two graph spaces, such as words embeddings, are isomorphic if they contain the same number of graph (words) vertices (or for a relaxed version, only for the most frequent k words) connected in the same way. Under this assumption,

[1] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain, email: magdalena.biesialska@upc.edu

[2] According to the results of a comparative study of cross-lingual embedding models presented in [13].

there have been many works [4, 8, 21] successfully utilizing orthogonal mapping methods to extract bilingual lexicons using CLEs. Importantly, such orthogonal transformations preserve length and inner products of vector representations of words.

The effectiveness of these orthogonal transformations falls drastically when the isomorphic condition does not hold. This problem can be observed for morphologically-rich or distant languages such as English and Japanese [25, 12], which in case of CLEs may be considered non-isomorphic language pairs [29]. Therefore, in this work, we report results for morphologically-rich languages such as German and Finnish (see Table 1).

2.2 Cross-Lingual Mapping Methods

This study only focuses on the methods based on word-level alignment. The vast majority of CLE models can be classified as **projection-based methods** (also referred as mapping-based) [19]. In this approach, word embeddings in two languages are trained independently on monolingual corpora, and next these word representations are mapped to a shared space using a linear transformation. The transformation matrix is usually learned from parallel data, such as word alignments or bilingual dictionaries [22]. Clearly, this method is suitable for supervised and weakly-supervised settings. Nevertheless, recently unsupervised models have made a successful breakthrough, showing that monolingual corpora alone are sufficient for learning the transformation.

Other CLE models fall into two categories: **pseudo-mixing methods** and **joint approaches** [22]. While the latter category of methods jointly optimizes monolingual and cross-lingual objectives, the former group does not rely on finding the mapping between the source and target language. Concretely, pseudo-mixing methods use word-level alignment from a seed dictionary to build a pseudo-bilingual corpus with source words being randomly replaced with their translations.

Since our proposed model leverages the projection-based approach, we will not further discuss the other two aforementioned methods, but rather we will concentrate on the mapping-based approaches. Existing projection-based methods can be classified into four groups:

- **regression methods** map the source language embeddings to the target language space using a least-squares objective [19, 9, 23];
- **canonical methods** map the word representations in both languages to a new shared space using canonical correlation analysis [11, 16, 2];
- **orthogonal methods** map the source language embeddings to maximize the similarity with the target language representations under the constraint of orthogonal transformation [27, 3, 24, 30];
- **margin methods** map the source language embeddings to maximize the margin between the correct translations and other candidates [15].

In fact, [6] demonstrated that regression, canonical and orthogonal methods can constitute a multi-step linear transformation framework.

3 METHODOLOGY

Our approach is motivated by the success of unsupervised CLE models, as well as recent promising results demonstrated by the refinement methods applied to supervised CLE models [10, 29]. In this work, we build upon these models; however, in contrast to the existing refinement approaches, we have designed our method to perform well in a more challenging unsupervised scenario.

The proposed method is based on the state-of-the-art approach introduced in [6], and extends it by applying additional transformations to refine the CLEs. In that respect, we follow the idea of [10] to create a cross-lingual vector space, which corresponds to the average of the aligned source and target language spaces. However, their method is supervised, and thus uses bilingual lexicon when performing a refinement of the initial alignment. Our method, on the other hand, is fully unsupervised and leverages self-learned seed dictionary to map the source and target language embeddings onto their average. More concretely, the training process is composed of three steps.

Firstly, having monolingual corpora for both source and target languages, word representations are learned for each language independently. In this step, word embedding methods such as Word2Vec [18], GloVe [20] or fastText [7] can be applied to obtain the monolingual embeddings.

Secondly, the source and target language embeddings are mapped to a shared vector space by means of a linear transformation. However, following [6], the vectors are normalized before the transformation is performed. This step normalizes the length of word vectors and performs dimension-wise mean centering. After this preprocessing step, embeddings can be aligned. While there exists a number of mapping methods (as described in Section 2.2), we will only explain approaches used in our model. At the outset, an initial seed lexicon needs to be learned in a fully unsupervised way. Artetxe *et al.* [6] employ a heuristic initialization method grounded in the idea that words in different languages have similar distributions, assuming that the embedding spaces are perfectly isomorphic (this is a simplification and in our proposed model we aim to fix it). Afterwards, the initial seed lexicon is improved using refinement methods.

Finally, in the proposed model, the last step is a self-supervised refinement of the alignment that is applied after the initial mapping is done. In general, the proposed method is motivated by the assumption that vector spaces of source and target language embeddings have different structure and should not be considered entirely isomorphic. Hence, when we operate in a shared cross-lingual space it is evident that source embeddings and their translations are still distant. Therefore, our refinement method consists of two phases.

Averaging the vectors. The underlying idea behind this step is to bring closer source words and their translations. Hence, for each word that is included in the induced dictionary, following the approach of [10], we shift each embedding vector to reach the middle point between the source word and its translation. More specifically, the vector average is computed in a standard way:

$$\vec{\mu}_{w,w'} = \frac{\vec{v}_w + \vec{v}_{w'}}{2} \tag{1}$$

where, $w \in V$ and $w' \in V'$ are source and target language words, and then the value is assigned to each embedding. To this end, we do not use any supervised source of parallel data, as the bilingual dictionary $D = \{(w, w')\}$ is induced during the initial alignment step.

Length normalization and mean centering. As the entire projection-based unsupervised CLE method relies on the orthogonal assumption; therefore, we concur with [29] that word embeddings should be of the same unit length. Moreover, they stress the importance of source and target language vector spaces having equal

Table 1. Bilingual Lexicon Induction results. Precision at k=1 (P@k x 100%) performance for Spanish (ES), German (DE) and Finnish (FI), where English (EN) is a source language.

	EN-ES		**EN-DE**		**EN-FI**	
	P@1	Δ	P@1	Δ	P@1	Δ
VecMap	37.47		48.47		33.08	
unsup. IterNorm + VecMap	36.33	-1.14	48.47	0.00	32.79	-0.29
Our method	**37.67**	**+0.20**	48.47	0.00	**33.29**	**+0.21**

magnitude centers. Therefore, every source and target word vector is transformed iteratively to fulfil both conditions:

$$\mathbf{y}_i^{(k)} = \mathbf{x}_i^{(k-1)} / \left\| \mathbf{x}_i^{(k-1)} \right\|_2 \tag{2}$$

and

$$\mathbf{x}_i^{(k)} = \mathbf{y}_i^{(k)} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i^{(k)} \tag{3}$$

respectively, where $x_i \in \{(\vec{v}_w, \vec{v}_{w'})\}$, $\|x_i\|_2 = 1$ for all i, and $\sum_{i=1}^{n} x_i = 0$.

4 EXPERIMENTS

In this section, we evaluate the quality of the CLE models on a standard task of bilingual lexicon induction.

4.1 Experimental Setup

Data. We conduct our experiments using a popular dataset introduced by Dinu *et al.* [9] and its extensions [4, 5]. The used dataset consists of the following language pairs: English-Spanish, English-German and English-Finnish. Monolingual embeddings of 300 dimensions were created using Word2Vec[3] [18] and were trained on WMT News Crawl (Spanish), WacKy crawling corpora (English, German), and Common Crawl (Finnish). To evaluate the performance, we use bilingual dictionaries provided in VecMap[4], where each test set consists of 1500 entries.

Baselines. We report our results in comparison with VecMap in the unsupervised mode [6]. Furthermore, we compare our results with a recent refinement method IterNorm[5] [29] which, contrary to the original paper, is used here in the unsupervised setting. We perform the evaluation using VecMap scripts with CSLS method used for retrieval (instead of nearest neighbor).

4.2 Bilingual Lexicon Induction

The intrinsic task of bilingual lexicon induction is a common choice to evaluate CLE models. The goal of this task is to indicate the most appropriate translation for each source word, given nearest neighbor target embeddings in the shared vector space. The accuracy is measured as the percentage of correctly translated source words with respect to a ground truth translation from a dictionary. This task is considered a good proxy for evaluating the performance of CLEs, as high-quality bilingual lexicons are available for many language pairs. However, one may argue that existing dictionaries merely contain the most frequent words and bilingual lexicon induction task should be accompanied by other evaluation methods [13] .

We follow the standard evaluation procedure by measuring scores for Precision at 1 (P@1), which determines how many times one of the correct translations of a source word is retrieved as the nearest neighbor of the source word in the target language. We report our results in Table 1.

We observe that our method outperforms baseline models in two cases: English-Spanish and English-Finnish. Furthermore, it performs on a par with baselines for the English-German language pair. As it can be seen, our method obtains better scores than IterNorm in all cases but one. While the method proposed in [29] was originally trained in the supervised setting, it is universal and can be applied in the case of an unsupervised CLE model as well. Although it achieved very good results in the supervised setting, according to our experiments, it does not perform as good when combined with the unsupervised VecMap model. We hypothesize, that the reason why our method surpasses the baselines is mainly due to the use of self-learned dictionary, which improves subsequent transformations.

5 CONCLUSION AND FUTURE WORK

This work adds to the growing body of research in CLEs. First, we introduced a self-supervised method to refine unsupervised bilingual word embeddings by leveraging a small self-learned seed lexicon. To our knowledge, this was the first attempt to apply a self-supervised refinement method to the state-of-the-art unsupervised CLE model by Artetxe *et al.* [6]. Second, our work addressed the problem of imperfect isomorphism in embedding vector spaces. The results, achieved in a bilingual dictionary induction task, suggest that our proposed approach improved the state-of-the-art for almost all evaluated language pairs.

In the future we plan to investigate if our method boosts the performance of existing models in downstream tasks, especially in unsupervised neural machine translation. Moreover, it would be also interesting to experiment with adapting our refinement technique to a multilingual alignment setting to improve cross-lingual transfer. In addition, as traditional (context-invariant) word embeddings suffer from the meaning conflation deficiency, a study of cross-lingual embeddings in relation to unsupervised sense representations and contextual embeddings would be interesting to perform.

ACKNOWLEDGEMENTS

We thank anonymous reviewers for their helpful comments. This work is supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development

[3] https://code.google.com/archive/p/word2vec/

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

[5] We would like to thank the authors for sharing with us a code snippet with an implementation of their method.

Fund and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, the contract TEC2015-69266-P (MINECO/FEDER,EU) and the contract PCIN-2017-079 (AEI/MINECO).

REFERENCES

[1] David Alvarez-Melis and Tommi Jaakkola, 'Gromov-Wasserstein alignment of word embedding spaces', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1881–1890, Brussels, Belgium, (October-November 2018). Association for Computational Linguistics.

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

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, 'Learning principled bilingual mappings of word embeddings while preserving monolingual invariance', in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294, (2016).

[4] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, 'Learning bilingual word embeddings with (almost) no bilingual data', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462, (2017).

[5] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, 'Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations', in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5012–5019, (2018).

[6] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, 'A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798, (2018).

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, 'Enriching word vectors with subword information', *Transactions of the Association for Computational Linguistics*, **5**, 135–146, (2017).

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

[9] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni, 'Improving zero-shot learning by mitigating the hubness problem', *Proceedings of ICLR*, (2015).

[10] Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert, 'Improving cross-lingual word embeddings by meeting in the middle', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 294–304, (2018).

[11] Manaal Faruqui and Chris Dyer, 'Improving vector space word representations using multilingual correlation', in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 462–471, Gothenburg, Sweden, (April 2014). Association for Computational Linguistics.

[12] Yoshinari Fujinuma, Jordan Boyd-Graber, and Michael J. Paul, 'A resource-free evaluation metric for cross-lingual word embeddings based on graph modularity', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4952–4962, Florence, Italy, (July 2019). Association for Computational Linguistics.

[13] Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić, 'How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 710–721, Florence, Italy, (July 2019). Association for Computational Linguistics.

[14] Edouard Grave, Armand Joulin, and Quentin Berthet, 'Unsupervised alignment of embeddings with wasserstein procrustes', in *AISTATS*, (2018).

[15] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni, 'Hubness and pollution: Delving into cross-space mapping for zero-shot learning', in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 270–280, Beijing, China, (July 2015). Association for Computational Linguistics.

[16] Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu, 'Deep multilingual correlation for improved word embeddings', in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 250–256, Denver, Colorado, (May–June 2015). Association for Computational Linguistics.

[17] Mike Maxwell and Baden Hughes, 'Frontiers in linguistic annotation for lower-density languages', in *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, LAC '06, pp. 29–37, Stroudsburg, PA, USA, (2006). Association for Computational Linguistics.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*, (2013).

[19] Tomas Mikolov, Quoc V Le, and Ilya Sutskever, 'Exploiting similarities among languages for machine translation', *arXiv preprint arXiv:1309.4168*, (2013).

[20] Jeffrey Pennington, Richard Socher, and Christopher Manning, 'Glove: Global vectors for word representation', in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, (2014).

[21] Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard, 'A discriminative latent-variable model for bilingual lexicon induction', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 458–468, Brussels, Belgium, (October-November 2018). Association for Computational Linguistics.

[22] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models, 2017.

[23] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto, 'Ridge regression, hubness, and zero-shot learning', *Lecture Notes in Computer Science*, 135–151, (2015).

[24] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla, 'Offline bilingual word vectors, orthogonal transformations and the inverted softmax', in *Proceedings of ICLR*, (2017).

[25] Anders Søgaard, Sebastian Ruder, and Ivan Vulić, 'On the limitations of unsupervised bilingual dictionary induction', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 778–788, (2018).

[26] Ivan Vulić and Anna Korhonen, 'On the role of seed lexicons in learning bilingual word embeddings', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 247–257, Berlin, Germany, (August 2016). Association for Computational Linguistics.

[27] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin, 'Normalized word embedding and orthogonal transform for bilingual word translation', in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1006–1011, Denver, Colorado, (May–June 2015). Association for Computational Linguistics.

[28] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun, 'Adversarial training for unsupervised bilingual lexicon induction', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1959–1970, Vancouver, Canada, (July 2017). Association for Computational Linguistics.

[29] Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber, 'Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3180–3189, Florence, Italy, (July 2019). Association for Computational Linguistics.

[30] Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola, 'Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings', in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1307–1317, San Diego, California, (June 2016). Association for Computational Linguistics.