Bilingual Lexicon Induction through Unsupervised Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Abstract

A recent research line has obtained strong results on bilingual lexicon induction by aligning independently trained word embeddings in two languages and using the resulting cross-lingual embeddings to induce word translation pairs through nearest neighbor or related retrieval methods. In this paper, we propose an alternative approach to this problem that builds on the recent work on unsupervised machine translation. This way, instead of directly inducing a bilingual lexicon from cross-lingual embeddings, we use them to build a phrase-table, combine it with a language model, and use the resulting machine translation system to generate a synthetic parallel corpus, from which we extract the bilingual lexicon using statistical word alignment techniques. As such, our method can work with any word embedding and cross-lingual mapping technique, and it does not require any additional resource besides the monolingual corpus used to train the embeddings. When evaluated on the exact same cross-lingual embeddings, our proposed method obtains an average improvement of 6 accuracy points over nearest neighbor and 4 points over CSLS retrieval, establishing a new state-of-the-art in the standard MUSE dataset.

1 Introduction

Cross-lingual word embedding mappings have attracted a lot of attention in recent times. These methods work by independently training word embeddings in different languages, and mapping them to a shared space through linear transformations. While early methods required a training dictionary to find the initial alignment (Mikolov et al., 2013), fully unsupervised methods have managed to obtain comparable results based on either adversarial training (Conneau et al., 2018) or self-learning (Artetxe et al., 2018b).

A prominent application of these methods is Bilingual Lexicon Induction (BLI), that is, using the resulting cross-lingual embeddings to build a bilingual dictionary. For that purpose, one would typically induce the translation of each source word by taking its corresponding nearest neighbor in the target language. However, it has been argued that this basic approach suffers from the hubness problem[1], which has motivated alternative retrieval methods like inverted nearest neighbor[2] (Dinu et al., 2015), inverted softmax (Smith et al., 2017), and Cross-domain Similarity Local Scaling (CSLS) (Conneau et al., 2018).

In this paper, we go one step further and, rather than directly inducing the bilingual dictionary from the cross-lingual word embeddings, we use them to build an unsupervised machine translation system, and extract a bilingual dictionary from a synthetic parallel corpus generated with it. This allows us to take advantage of a strong language model and naturally extract translation equivalences through statistical word alignment. At the same time, our method can be used as a drop-in replacement of traditional retrieval techniques, as it can work with any cross-lingual word embeddings and it does not require any additional resource besides the monolingual corpus used to train them. Our experiments show the effectiveness of this alternative approach, which outperforms the previous best retrieval method by 4 accuracy points on average, establishing a new state-of-the-art in the standard MUSE dataset. As such, we conclude that, contrary to recent trend, future research in BLI should not focus exclusively on direct retrieval methods.

[1] Hubness (Radovanović et al., 2010a,b) refers to the phenomenon of a few points being the nearest neighbors of many other points in high-dimensional spaces, which has been reported to severely affect cross-lingual embedding mappings (Dinu et al., 2015).

[2] The original paper refers to this method as *globally corrected* retrieval.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5002–5007
Florence, Italy, July 28 - August 2, 2019.

2 Proposed method

The input of our method is a set of cross-lingual word embeddings and the monolingual corpora used to train them. In our experiments, we use fastText embeddings (Bojanowski et al., 2017) mapped through VecMap (Artetxe et al., 2018b), but the algorithm described next can also work with any other word embedding and cross-lingual mapping method.

The general idea of our method is to to build an unsupervised phrase-based statistical machine translation system (Lample et al., 2018; Artetxe et al., 2018c, 2019), and use it to generate a synthetic parallel corpus from which to extract a bilingual dictionary. For that purpose, we first derive phrase embeddings from the input word embeddings by taking the 400,000 most frequent bigrams and and the 400,000 most frequent trigrams in each language, and assigning them the centroid of the words they contain. Having done that, we use the resulting cross-lingual phrase embeddings to build a phrase-table as described in Artetxe et al. (2018c). More concretely, we extract translation candidates by taking the 100 nearest-neighbors of each source phrase, and score them with the softmax function over their cosine similarities:

$$\phi(\bar{f}|\bar{e}) = \frac{\exp\left(\cos(\bar{e},\bar{f})/\tau\right)}{\sum_{\bar{f}'}\exp\left(\cos(\bar{e},\bar{f}')/\tau\right)}$$

where the temperature τ is estimated using maximum likelihood estimation over a dictionary induced in the reverse direction. In addition to the phrase translation probabilities in both directions, we also estimate the forward and reverse lexical weightings by aligning each word in the target phrase with the one in the source phrase most likely generating it, and taking the product of their respective translation probabilities.

We then combine this phrase-table with a distortion model and a 5-gram language model estimated in the target language corpus, which results in a phrase-based machine translation system. So as to optimize the weights of the resulting model, we use the unsupervised tuning procedure proposed by Artetxe et al. (2019), which combines a cyclic consistency loss and a language modeling loss over a subset of 2,000 sentences from each monolingual corpora.

Having done that, we generate a synthetic parallel corpus by translating the source language monolingual corpus with the resulting machine translation system.[3] We then word align this corpus using FastAlign (Dyer et al., 2013) with default hyperparameters and the *grow-diag-final-and* symmetrization heuristic. Finally, we build a phrase-table from the word aligned corpus, and extract a bilingual dictionary from it by discarding all non-unigram entries. For words with more than one entry, we rank translation candidates according to their direct translation probability.

3 Experimental settings

In order to compare our proposed method head-to-head with other BLI methods, the experimental setting needs to fix the monolingual embedding training method, as well as the cross-lingual mapping algorithm and the evaluation dictionaries. In addition, in order to avoid any advantage, our method should not see any further monolingual corpora than those used to train the monolingual embeddings. Unfortunately, existing BLI datasets distribute pre-trained word embeddings alone, but not the monolingual corpora used to train them. For that reason, we decide to use the evaluation dictionaries from the standard MUSE dataset (Conneau et al., 2018) but, instead of using the pre-trained Wikipedia embeddings distributed with it, we extract monolingual corpora from Wikipedia ourselves and train our own embeddings trying to be as faithful as possible to the original settings. This allows us to compare our proposed method to previous retrieval techniques in the exact same conditions, while keeping our results as comparable as possible to previous work reporting results for the MUSE dataset.

More concretely, we use WikiExtractor[4] to extract plain text from Wikipedia dumps, and preprocess the resulting corpus using standard Moses tools (Koehn et al., 2007) by applying sentence splitting, punctuation normalization, tokenization with aggressive hyphen splitting, and lowercasing. We then train word embeddings for each language using the skip-gram implementation of fastText (Bojanowski et al., 2017) with default hyperparameters, restricting the vocabulary to the 200,000 most frequent tokens. The official embeddings in

[3] For efficiency purposes, we restricted the size of the synthetic parallel corpus to a maximum of 10 million sentences, and use cube-pruning for faster decoding. As such, our results could likely be improved by translating the full monolingual corpus with standard decoding.

[4] https://github.com/attardi/wikiextractor

	en-es		en-fr		en-de		en-ru		avg.
	→	←	→	←	→	←	→	←	
Nearest neighbor	81.9	82.8	81.6	81.7	73.3	72.3	44.3	65.6	72.9
Inv. nearest neighbor (Dinu et al., 2015)	80.6	77.6	81.3	79.0	69.8	69.7	43.7	54.1	69.5
Inv. softmax (Smith et al., 2017)	81.7	82.7	81.7	81.7	73.5	72.3	44.4	65.5	72.9
CSLS (Conneau et al., 2018)	82.5	84.7	83.3	83.4	75.6	75.3	47.4	67.2	74.9
Proposed method	**87.0**	**87.9**	**86.0**	**86.2**	**81.9**	**80.2**	**50.4**	**71.3**	**78.9**

Table 1: P@1 of proposed system and previous retrieval methods, using the same cross-lingual embeddings.

the MUSE dataset were trained using these exact same settings, so our embeddings only differ in the Wikipedia dump used to extract the training corpus and the pre-processing applied to it, which is not documented in the original dataset.

Having done that, we map these word embeddings to a cross-lingual space using the unsupervised mode in VecMap (Artetxe et al., 2018b), which builds an initial solution based on the intra-lingual similarity distribution of the embeddings and iteratively improves it through self-learning. Finally, we induce a bilingual dictionary using our proposed method and evaluate it in comparison to previous retrieval methods (standard nearest neighbor, inverted nearest neighbor, inverted softmax[5] and CSLS). Following common practice, we use precision at 1 as our evaluation measure.[6]

4 Results and discussion

Table 1 reports the results of our proposed system in comparison to previous retrieval methods. As it can be seen, our method obtains the best results in all language pairs and directions, with an average improvement of 6 points over nearest neighbor and 4 points over CSLS, which is the best performing previous method. These results are very consistent across all translation directions, with an absolute improvement between 2.7 and 6.3 points over CSLS. Interestingly, neither inverted nearest neighbor nor inverted softmax are able to outperform standard nearest neighbor, presumably because our cross-lingual embeddings are less sensitive to hubness thanks to the symmetric re-weighting in VecMap (Artetxe et al., 2018a). At the same time, CSLS obtains an absolute improvement of 2 points over nearest neighbor, only a third of what our method achieves. This suggests that, while previous retrieval methods have almost exclusively focused on addressing the hubness problem, there is a substantial margin of improvement beyond this phenomenon.

So as to put these numbers into perspective, Table 2 compares our method to previous results reported in the literature.[7] As it can be seen, our proposed method obtains the best published results in all language pairs and directions, outperforming the previous state-of-the-art by a substantial margin. Note, moreover, that these previous systems mostly differ in their cross-lingual mapping algorithm and not the retrieval method, so our improvements are orthogonal.

We believe that, beyond the substantial gains in this particular task, our work has **important implications** for future research in cross-lingual word embedding mappings. While most work in this topic uses BLI as the only evaluation task, Glavas et al. (2019) recently showed that BLI results do not always correlate well with downstream performance. In particular, they observe that some mapping methods that are specifically designed for BLI perform poorly in other tasks. Our work shows that, besides their poor performance in those tasks, these BLI-centric mapping methods might not even be the optimal approach to BLI, as our alternative method, which relies on unsupervised machine translation instead of direct

[5]Inverted softmax has a temperature hyperparameter T, which is typically tuned in the training dictionary. Given that we do not have any training dictionary in our fully unsupervised settings, we use a fixed temperature of $T = 30$, which was also used by some previous authors (Lample et al., 2018). While we tried other values in our preliminary experiments, but we did not observe any significant difference.

[6]We find a few out-of-vocabularies in the evaluation dictionary that are likely caused by minor pre-processing differences. In those cases, we use copying as a back-off strategy (i.e. if a given word is not found in our induced dictionary, we simply leave it unchanged). In any case, the percentage of out-of-vocabularies is always below 1%, so this has a negligible effect in the reported results.

[7]Note that previous results are based on the pre-trained embeddings of the MUSE dataset, while we had to train our embeddings to have a controlled experiment (see Section 3). In any case, our embeddings are trained following the official dataset setting, using Wikipedia, the same system and hyperparameters, so our results should be roughly comparable.

	en-es		en-fr		en-de		en-ru		avg.
	→	←	→	←	→	←	→	←	
Conneau et al. (2018)	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	72.3
Hoshen and Wolf (2018)	82.1	84.1	82.3	82.9	74.7	73.0	47.5	61.8	73.6
Grave et al. (2018)	82.8	84.1	82.6	82.9	75.4	73.3	43.7	59.1	73.0
Alvarez-Melis and Jaakkola (2018)	81.7	80.4	81.3	78.9	71.9	72.8	45.1	43.7	69.5
Yang et al. (2018)	79.9	79.3	78.4	78.9	71.5	70.3	-	-	-
Mukherjee et al. (2018)	84.5	79.2	-	-	-	-	-	-	-
Alvarez-Melis et al. (2018)	81.3	81.8	82.9	81.6	73.8	71.1	41.7	55.4	71.2
Xu et al. (2018)	79.5	77.8	77.9	75.5	69.3	67.0	-	-	-
Proposed method	**87.0**	**87.9**	**86.0**	**86.2**	**81.9**	**80.2**	**50.4**	**71.3**	**78.9**

Table 2: Results of the proposed method in comparison to previous work (P@1). All systems are fully unsupervised and use fastText embeddings trained on Wikipedia with the same hyperparameters.

retrieval over mapped embeddings, obtains substantially better results without requiring any additional resource. As such, we argue that 1) future work in cross-lingual word embeddings should consider other evaluation tasks in addition to BLI, and 2) future work in BLI should consider other alternatives in addition to direct retrieval over cross-lingual embedding mappings.

5 Related work

While BLI has been previously tackled using count-based vector space models (Vulić and Moens, 2013) and statistical decipherment (Ravi and Knight, 2011; Dou and Knight, 2012), these methods have recently been superseded by cross-lingual embedding mappings, which work by aligning independently trained word embeddings in different languages. For that purpose, early methods required a training dictionary, which was used to learn a linear transformation that mapped these embeddings into a shared cross-lingual space (Mikolov et al., 2013; Artetxe et al., 2018a). The resulting cross-lingual embeddings are then used to induce the translations of words that were missing in the training dictionary by taking their nearest neighbor in the target language.

The amount of required supervision was later reduced through self-learning methods (Artetxe et al., 2017), and then completely eliminated through adversarial training (Zhang et al., 2017a; Conneau et al., 2018) or more robust iterative approaches combined with initialization heuristics (Artetxe et al., 2018b; Hoshen and Wolf, 2018). At the same time, several recent methods have formulated embedding mappings as an optimal transport problem (Zhang et al., 2017b; Grave et al., 2018; Alvarez-Melis and Jaakkola, 2018).

In addition to that, a large body of work has focused on addressing the hubness problem that arises when directly inducing bilingual dictionaries from cross-lingual embeddings, either through the retrieval method (Dinu et al., 2015; Smith et al., 2017; Conneau et al., 2018) or the mapping itself (Lazaridou et al., 2015; Shigeto et al., 2015; Joulin et al., 2018). While all these previous methods directly induce bilingual dictionaries from cross-lingually mapped embeddings, our proposed method combines them with unsupervised machine translation techniques, outperforming them all by a substantial margin.

6 Conclusions and future work

We propose a new approach to BLI which, instead of directly inducing bilingual dictionaries from cross-lingual embedding mappings, uses them to build an unsupervised machine translation system, which is then used to generate a synthetic parallel corpus from which to extract bilingual lexica. Our approach does not require any additional resource besides the monolingual corpora used to train the embeddings, and outperforms traditional retrieval techniques by a substantial margin. We thus conclude that, contrary to recent trend, future work in BLI should not focus exclusively in direct retrieval approaches, nor should BLI be the only evaluation task for cross-lingual embeddings. Our code is available at `https://github.com/artetxem/monoses`.

In the future, we would like to further improve our method by incorporating additional ideas from unsupervised machine translation such as joint refinement and neural hybridization (Artetxe et al., 2019). In addition to that, we would like to integrate our induced dictionaries in other downstream

tasks like unsupervised cross-lingual information retrieval (Litschko et al., 2018).

Acknowledgments

This research was partially supported by the Spanish MINECO (UnsupNMT TIN2017-91692-EXP and DOMINO PGC2018-102041-B-I00, co-funded by EU FEDER), the BigKnowledge project (BBVA foundation grant 2018), the UPV/EHU (excellence research group), and the NVIDIA GPU grant program. Mikel Artetxe was supported by a doctoral grant from the Spanish MECD.

References

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.

David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. 2018. Towards optimal transport with global invariances. *arXiv preprint arXiv:1806.09277*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. 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)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5012–5019.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018c. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), workshop track*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, Jeju Island, Korea. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508*.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with wasserstein procrustes. *arXiv preprint arXiv:1805.11222*.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Herve Jegou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of*

the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. 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)*, pages 270–280. Association for Computational Linguistics.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. *arXiv preprint arXiv:1805.00879*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tanmoy Mukherjee, Makoto Yamada, and Timothy Hospedales. 2018. Learning unsupervised word translations without adversaries. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 627–632, Brussels, Belgium. Association for Computational Linguistics.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010a. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010b. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193. ACM.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. 2015. *Ridge Regression, Hubness, and Zero-Shot Learning*, pages 135–151. Springer International Publishing.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations (ICLR 2017)*.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1613–1624, Seattle, Washington, USA. Association for Computational Linguistics.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng Yang, Fuli Luo, Shuangzhi Wu, Jingjing Xu, Dongdong Zhang, and Xu Sun. 2018. Learning unsupervised word mapping by maximizing mean discrepancy. *arXiv preprint arXiv:1811.00275*.

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

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.