Questioning the Use of Bilingual Lexicon Induction as an Evaluation Task for Bilingual Word Embeddings

Benjamin Marie **Atsushi Fujita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{bmarie, atsushi.fujita}@nict.go.jp

1 Introduction

Word embeddings have been shown useful in a wide range of natural language processing tasks. Recently, many methods have been proposed to transform monolingual word embeddings into bilingual word embeddings (BWE), relying on a small seed bilingual lexicon as a weak supervision (Mikolov et al., 2013). More recently, unsupervised methods, i.e., that do not use bilingual resources for training, have also been proposed (Artetxe et al., 2017; Lample et al., 2018).

Most of the methods for learning BWE are evaluated through bilingual lexicon induction (BLI) tasks, where a test set comprises 1k–3k source words each paired with one or several correct word translations in the target language. For each source word, a translation is retrieved from the target language vocabulary relying only on the BWE to be evaluated. Then, the accuracy of the retrieval, i.e., the percentage of source words for which one of the correct word translations has been retrieved, is computed as a quality measure of the BWE. Previous work on BWE has claimed the superiority of a method for learning BWE if its accuracy is higher than other evaluated methods. BLI remains the most used evaluation task since the work by Mikolov et al. (2013), mainly because of its low computational cost and the relatively low cost of creating many test sets for several language pairs.

In this paper, we focus on assessing how general are the conclusions drawn from BLI-based evaluations of BWE. The purpose of this paper is to shed light on the limits of current practices rather than proposing alternative evaluation methods or finding which BWE performs the best. We re-evaluate nine methods for learning BWE through BLI tasks with various experimental settings varying in the test set, the monolingual word embeddings to transform, and the method for retrieving word translations. These are independent of the methods used for training BWE. In other words, the methods and their parameters remain unchanged across our experiments. Through our experiments to highlight the significant limits of evaluating BWE on BLI tasks, we identify several incorrect practices, such as the use of inconsistent evaluation settings. One of our most surprising findings is that the seminal work by Mikolov et al. (2013) still remains a very competitive method for BLI.

2 Methods Evaluated

We sampled a small number of popular methods for learning BWE to be re-evaluated on BLI tasks. Note that since we aim at analyzing common practices in evaluating BWE, our conclusions would hold for other BWE methods.

2.1 Weakly-supervised BWE

`Miko` (Mikolov et al., 2013): a projection matrix is trained from source word embeddings to the target embedding space. This method requires the lowest computational cost for training among the methods we re-evaluate. It is also often considered as a weak baseline method for BLI tasks, since it has been shown to underperform more recent methods.

`Miko-c` (Artetxe et al., 2016): a modified version of `Miko` that pre-processes the word embeddings with mean centering and length normalization before projection.

`VM16` (Artetxe et al., 2016): similar to `Miko-c` but the projection is performed through orthogonal mapping as proposed by Xing et al. (2015). This method has been shown to outperform `Miko`.

`VM17-S` (Artetxe et al., 2017): significantly different from `Miko`, since it relies on a self-learning algorithm that first induces a poorly accurate bilingual lexicon and refines it iteratively. For initialization of the training, we used the same training dictionaries used for training other weakly-supervised methods.

`VM18-S` (Artetxe et al., 2018a): a generalization of previous work that uses a multi-step framework to pre-process and post-process projected word em-

beddings. This method has been shown to outperform previous work in BLI tasks.

Muse-S (Lample et al., 2018): similar to VM17-S but uses a different algorithm for inducing the bilingual lexicon in the refinement steps.

2.2 Unsupervised BWE

VM17-U (Artetxe et al., 2017): similar to VM17-S but uses a bilingual dictionary made of pairs of numbers for initialization.

Muse-U (Lample et al., 2018): performs first adversarial training to compute the mapping of BWE and generate a bilingual lexicon that is then used to train Muse-S and refine the mapping. This method outperforms previous work but has been shown to be highly unstable by subsequent work (Artetxe et al., 2018b; Søgaard et al., 2018).

VM18-U (Artetxe et al., 2018b): an extension of VM18-S and VM17-S for unsupervised BWE. It has been shown to consistently outperform previous work, even weakly-supervised BWE.

3 Datasets and Tools

We re-evaluated all the methods with publicly available frameworks and datasets for English-to-German (en-de), English-to-Spanish (en-es), English-to-Italian (en-it), and English-to-Finnish (en-fi). For each language, we used two sets of word embeddings for 200k words (300 dimensions): **Wikipedia-emb**, trained with fastText on the Wikipedia data of the corresponding language,[1] and **Vecmap-emb**, trained using word2vec on diverse corpora.[2,3] We expect BWE trained using Wikipedia-emb to be much more accurate than with Vecmap-emb, since Wikipedia corpora are comparable to some degree across languages, while Vecmap-emb has been trained on much more diverse sets of monolingual corpora.

As test bilingual lexicon for evaluation, we used four **Muse test sets**,[4,5] containing only frequent source words of Wikipedia, and four **Vecmap test sets**, containing source words of various frequency. To avoid any overlap between the test data and the training data of the weakly-supervised methods, we used Muse training data when performing an evaluation on Muse test sets and Vecmap training data for an evaluation on Vecmap test sets. These training data, also provided with Muse and Vecmap test sets, contain 5,000 word pairs for very frequent source words.

For contrastive experiments with a non-European language pair, we added English-to-Japanese (en-ja). Word embeddings were trained as for Vecmap-emb on the NTCIR monolingual data in patent domain (Goto et al., 2013). The training and testing bilingual lexicons were created by ourselves following the same procedure proposed by Dinu et al. (2014).[6] Even though Japanese and English are distant languages, we can expect a reasonable accuracy of BWE for this BLI task since all the used data are from the same controlled domain.

To train BWE, we used the Vecmap toolkit for Miko, Miko-c, VM16, VM17-S, VM17-U, VM18-S, and VM18-U, and the Muse toolkit for Muse-S and Muse-U.

To retrieve the best translation in the target vocabulary of 200k words, we used the Vecmap toolkit with nearest neighbors (NN) or cross-domain similarity scaling (CSLS) (Lample et al., 2018).

4 Impact of Using Consistent Settings

The left half of Table 1 shows the results on all the test sets obtained with Vecmap-emb and CSLS. Since each method had been proposed to outperform the other pre-existing methods, we grouped the methods in the table according to their (pre-)publication dates. Methods in a lower group in the table should thus outperform those presented in the upper groups.

First of all, none of the evaluated methods consistently worked best for all the test sets. More importantly, previously proposed methods, such as Miko and Miko-c, outperformed more recent methods for many test sets. This happens because different methods have so far been compared in combination with different translation retrieval methods, such as NN, inverted softmax (Smith et al., 2017), or CSLS. For instance, Artetxe et al. (2018b) concluded that VM18-U surpasses VM18-S, on the basis of a comparison between VM18-S's accuracy with inverted softmax and VM18-U's accuracy with CSLS. In contrast, through

[1] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

[2] Wacky corpora (English, Italian, German), Common Crawl corpus (Finnish), and News Crawl corpus (Spanish).

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

[4] We use the official Muse test sets.

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

[6] This procedure was also used to create Vecmap bilingual lexicons.

Method	CSLS									NN								
	Muse test set				Vecmap test set				NTCIR	Muse test set				Vecmap test set				NTCIR
	en-de	en-es	en-fi	en-it	en-de	en-es	en-fi	en-it	en-ja	en-de	en-es	en-fi	en-it	en-de	en-es	en-fi	en-it	en-ja
Miko	66.35	54.23	54.86	56.97	44.00	35.80	32.87	45.60	**57.00**	59.31	46.04	48.77	39.75	35.00	27.73	25.91	34.93	52.60
Miko-c	**66.49**	54.23	**55.89**	59.85	43.47	35.33	33.64	46.53	55.67	59.44	45.03	49.11	48.99	35.07	26.47	25.91	38.47	51.47
VM16	65.13	54.63	52.53	58.04	47.27	35.73	33.92	45.07	54.93	63.71	51.21	**50.82**	51.41	41.87	31.40	30.62	39.27	55.27
VM17-S	60.93	48.32	44.38	53.22	47.13	33.07	30.76	43.87	55.53	59.44	45.37	43.84	47.72	40.80	28.07	28.86	39.33	54.53
VM17-U	60.05	46.98	45.07	53.89	46.60	31.60	29.85	44.33	53.67	59.38	43.09	44.45	47.65	40.73	28.47	26.33	38.47	54.40
Muse-S	62.69	54.23	48.63	56.90	46.80	36.47	31.88	45.33	54.40	62.15	50.60	48.97	51.47	40.93	31.00	28.37	39.60	55.20
Muse-U	0.00	0.00	0.00	57.04	0.00	0.00	0.00	44.93	54.27	0.00	0.00	0.00	51.88	0.00	0.00	0.00	39.27	**55.43**
VM18-S	63.98	55.77	50.48	**63.34**	47.20	**38.20**	**34.97**	47.33	51.27	**64.45**	**56.17**	50.62	**58.85**	**44.27**	**36.47**	**32.79**	**43.80**	54.07
VM18-U	64.45	**56.17**	47.19	59.25	**48.47**	37.40	32.94	**48.27**	52.87	63.44	52.75	45.62	55.43	43.87	34.00	29.92	43.53	54.67

Table 1: Accuracy using Vecmap-emb. **Bold** indicates the best score in each column.

a consistent use of CSLS for all the evaluated methods, we conclude that weakly-supervised BWE still remains competitive to or better than unsupervised BWE in BLI. For instance, we observed differences of more than 8 and 4 points of accuracy for Muse en-fi and Muse en-it test sets, respectively. These results point out that, for *a fair comparison of BWE* through BLI tasks, we must use the same translation retrieval method for all the evaluated BWE to draw meaningful conclusions.

Muse-U achieved an accuracy of 0.00 in many tasks, whereas Lample et al. (2018) did not observe any 0.00 accuracy thanks to Wikipedia-emb. It is much harder to learn BWE with Muse-U when using embeddings trained on monolingual corpora from diverse domains such as Vecmap-emb (see Section 6).

5 Impact of the Retrieval Methods

The accuracy in BLI tasks can be dramatically improved depending on the method for retrieving the word translations, given some BWE. For instance, inverted softmax and CSLS have both been shown to provide a better accuracy than NN. Given that the retrieval method has no impact on the embeddings quality itself, we should separately evaluate BWE and the retrieval method for BLI.

In Table 1, we compare the accuracy obtained with CSLS (left) and NN (right). NN leads to lower accuracy than CSLS for most of the tasks and the BWE methods. Especially, while Miko appeared to be a competitive BWE with CSLS, its accuracy dropped by more than 4 points for all tasks with NN. A more notable finding is that the best BWE (and the ranking of BWE methods) for each task is not consistent between CSLS and NN. With NN, VM18-S performed the best for 7 out of 9 tasks.

NN and CSLS rely on different aspects of learned BWE, and from the above observations, we conclude that the performance of some BWE in BLI tasks can be compared only with the same word translation retrieval method.

Method	Muse test set				Vecmap test set			
	en-de	en-es	en-fi	en-it	en-de	en-es	en-fi	en-it
Miko	6.27	26.53	-6.20	18.80	8.60	24.00	-5.14	12.00
Miko-c	8.13	27.60	-2.73	17.94	10.27	25.14	-2.67	12.20
VM16	10.60	26.87	-2.13	18.94	6.00	24.34	-5.76	14.20
VM17-S	12.87	33.33	-2.33	23.19	5.87	27.53	-5.27	16.73
Muse-S	12.99	27.66	-0.86	20.67	7.14	23.93	-2.39	14.28
Muse-U	74.87	82.13	43.13	20.99	54.93	60.07	28.23	14.34
VM18-S	10.53	25.13	-0.20	14.13	6.47	22.00	-3.58	10.47
VM18-U	11.67	26.48	3.67	19.86	6.53	24.20	-2.61	12.46

Table 2: Difference of corrected accuracy obtained with Wikipedia-emb and Vecmap-emb through CSLS.

6 Impact of the Monolingual Word Embeddings

Lample et al. (2018) reported much higher accuracies when using Wikipedia-emb, for Muse test sets, than the accuracies observed in Sections 4 and 5. We assume that Wikipedia-emb makes the BLI tasks easier and report in Table 2 on the relative changes of the corrected accuracy[7] when using Wikipedia-emb instead of Vecmap-emb.

BWE trained on Wikipedia-emb largely outperformed BWE trained on Vecmap-emb. For instance, we obtained more than 20 accuracy points of improvements for en-es test sets, and an absolute accuracy above 60.00 for all test sets of en-de, en-es, and en-it.

[7]The percentage of source words in a test set whose correct translation is found by a particular setting. Since the Wikipedia-emb and Vecmap-emb vocabularies partially cover the word pairs in the Vecmap and Muse test sets, respectively, we compute the coverage c of the given BWE, as defined in the Vecmap toolkit, and the corrected accuracy by multiplying the accuracy by c. We do not provide the results for VM17-U, since Wikipedia-emb does not provide embeddings for numerals.

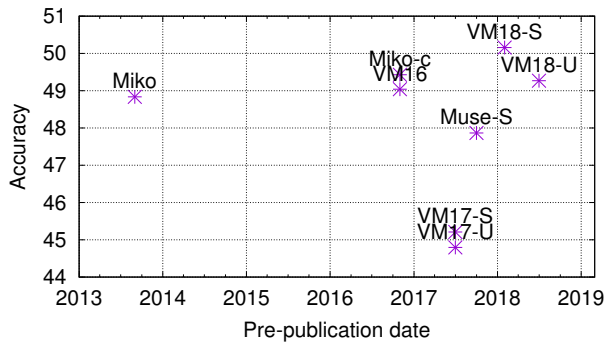

Figure 1: Average accuracy across eight public test sets achieved by each method with Vecmap-emb and CSLS. `Muse-U` is located far below.

As reported in Lample et al. (2018), `Muse-U` becomes competitive when trained on Wikipedia-emb.

One exception appears for en-fi test sets: using Wikipedia-emb led to a lower accuracy for most of our evaluated BWE. We assume that the Finnish Wikipedia corpus was not large enough to train accurate monolingual word embeddings. In fact, this corpus was significantly smaller than the Wikipedia corpora for all the other languages, and than the Finnish Common Crawl corpus used to train Finnish Vecmap-emb.

Another finding is that Wikipedia-emb helps unsupervised methods, i.e., `Muse-U` and `VM18-U`, much more than their weakly-supervised counterparts, i.e., `Muse-S` and `VM18-S`. Furthermore, with Wikipedia-emb, we have yet another conclusion, i.e., unsupervised BWE are better than weakly-supervised BWE for most of the tasks, in contradiction to what we concluded from Vecmap-emb (Section 4). As shown by Søgaard et al. (2018), a possible explanation is that the comparability of Wikipedia corpora facilitates unsupervised training.

7 Summary

We demonstrated that conclusions drawn by previous work are not valid in various settings according to BLI tasks. In other words, the superiority of some BWE can only be demonstrated for particular BLI experimental settings and not in general as in done by previous work. Moreover, even though recent methods for BWE are optimized for BLI, using CSLS during training (Artetxe et al., 2018b; Lample et al., 2018), potentially overfitting to the BLI task, it remains unclear whether any significant progress has really been achieved in BWE for BLI in the past 5 years, as illustrated by Figure 1. Moreover, a concurrent work of Glavaš et al. (2019) also shows that we need to re-assess existing baselines and design new evaluation protocols for BWE.

References

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*.

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 AAAI*.

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 ACL*.

Georgiana Dinu, Angeliki Lazaridou, and Marco G Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *CoRR*, abs/1412.6568.

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. pages 710–721.

Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. 2013. Overview of the patent machine translation task at the NTCIR-10 workshop. In *Proceedings of NTCIR Conference*.

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

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

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*.

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*.