Unsupervised Joint Training of 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

Abstract

State-of-the-art methods for unsupervised bilingual word embeddings (BWE) train a mapping function that maps pre-trained monolingual word embeddings into a bilingual space. Despite its remarkable results, unsupervised mapping is also well-known to be limited by the dissimilarity between the original word embedding spaces to be mapped. In this work, we propose a new approach that trains unsupervised BWE jointly on synthetic parallel data generated through unsupervised machine translation. We demonstrate that existing algorithms that jointly train BWE are very robust to noisy training data and show that unsupervised BWE jointly trained significantly outperform unsupervised mapped BWE in several cross-lingual NLP tasks.

1 Introduction

Bilingual word embeddings (BWE) represent the vocabulary of two languages in one common continuous vector space. They are known to be useful in a wide range of cross-lingual NLP tasks.

The most prevalent methods for training BWE are so-called mapping methods (Mikolov et al., 2013a): word embeddings for two languages are separately trained on respective monolingual data and then mapped into one common embedding space. The mapping function is usually trained using a small bilingual lexicon for supervision. Recently, unsupervised mapping for BWE (Artetxe et al., 2018a; Lample et al., 2018a), i.e., trained without using any manually created bilingual resources, has been shown to reach a performance comparable to supervised BWE in several cross-lingual NLP tasks. Unsupervised BWE are trained with a three-step approach. First, word embeddings are roughly mapped into an initial BWE space, for instance using adversarial training or an heuristic mapping. Then, using the initial BWE, a small synthetic bilingual lexicon is induced. Finally, a new BWE, which is expected to be better than the initial BWE, is learned from the induced lexicon through a pseudo-supervision with some supervised mapping method. The last two steps can be repeated to refine the BWE.

In spite of their success, unsupervised mapping methods are inherently limited by the dissimilarity between the original word embedding spaces to be mapped. The feasibility of aligning two embedding spaces relies on the assumption that they are isomorphic. However, Søgaard et al. (2018) showed that these spaces are, in general, far from being isomorphic, and thus they result in sub-optimal or degenerated unsupervised mappings.

On the other hand, supervised methods that jointly train BWE from scratch (Upadhyay et al., 2016), on parallel or comparable corpora, do not have such limits since no pre-existing embedding spaces and no mapping function are involved. These methods jointly train BWE by exploiting bilingual and monolingual contexts of words, materialized by sentence or document pairs, to learn a single BWE space. However, they require large bilingual resources for training. To the best of our knowledge, joint training of BWE has never been explored for unsupervised scenarios.

In this paper, we propose *unsupervised joint training of BWE*. Our method is an extension of previous work on unsupervised BWE: we propose to generate, without supervision, synthetic parallel sentences that can be directly exploited to jointly train BWE with existing algorithms. We empirically show that this method learns better BWE for several cross-lingual NLP tasks.

2 Pseudo-supervised joint training

On the strong assumption that existing algorithms for joint training of BWE are robust enough even

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3224–3230
Florence, Italy, July 28 - August 2, 2019.

with very noisy parallel training data, we formulate the following research question:

> *Do synthetic sentence pairs supply useful bilingual contextual information for learning better BWE?*

2.1 Bilingual skipgram

Previous work on joint training of BWE hypothesizes that exploiting both monolingual and bilingual contextual information yields better word embeddings, monolingually and bilingually.

Among several existing algorithms for joint training of BWE, in this work, we use bilingual skipgram (BIVEC) (Luong et al., 2015), which has been shown to outperform other methods in several NLP tasks (Upadhyay et al., 2016). BIVEC uses the skipgram algorithm (Mikolov et al., 2013b) to learn the word embeddings for each language and exploits word alignments obtained for parallel data in order to make the embeddings cross-lingual. Given a pair of sentences, S_1 in some language L1 and S_2 in another language L2, a word w_i in S_1 is replaced with its aligned word $a(w_i)$ in S_2, so that the L1 context can also be used for learning the embedding of the L2 word. BIVEC has been shown to be robust to noisy word alignments (Luong et al., 2015), which is a significant advantage of this method in our scenario using synthetic parallel data.

2.2 Training on synthetic parallel data

For an unsupervised training of BWE, the training data must also be generated in an unsupervised way. To this end, we chose unsupervised machine translation (MT). Recent work has shown significant progress in unsupervised MT (Artetxe et al., 2018b; Lample et al., 2018b) with generated translations of a reasonable quality. Both statistical (SMT) and neural MT (NMT) have been adapted to the unsupervised scenario. We chose unsupervised SMT (USMT) to generate synthetic parallel data since it generates better translations than unsupervised NMT (Lample et al., 2018b).

Given an initial BWE, for instance learned with unsupervised mapping methods, our method works as follows (see also Figure 1). First, a USMT is trained from monolingual data. We collect a set of phrases made of up to L tokens, using word2phrase,[1] for each of the source and target languages. As phrases, we also consider all the token types in each corpus. In our phrase table, each L1 phrase is paired with its k most probable translations in L2 determined based on a score computed from the given BWE.[2] The phrase table and a language model trained on the L2 monolingual data compose the initial USMT. Then, the USMT is iteratively refined in the following manner.

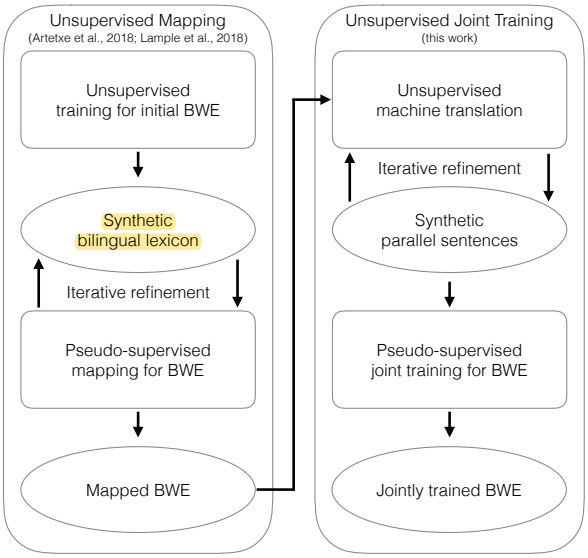

Figure 1: Our joint training framework is on top of existing unsupervised mapping methods.

- Synthetic parallel data are generated by translating monolingual data using the USMT. Both L1-to-L2 and L2-to-L1 translations can be considered (Artetxe et al., 2019).
- A new phrase table is trained on the synthetic parallel data to form a new USMT.

Finally, on the synthetic parallel data generated by our USMT after N refinement steps, we jointly train new BWE as described in Section 2.1.

Although this approach can efficiently generate parallel data of a reasonable quality, as shown in Figure 1, it heavily relies on the feasibility of mapping the word embeddings learned for L1 and L2 in the same space and used for the initial USMT. If the mapping fails, we cannot expect USMT to generate useful data for jointly training BWE. Conversely, if the mapping succeeds, we can generate data with bilingual contexts that may be useful to jointly train BWE.

More importantly, we use USMT assuming that BIVEC is robust enough to learn from very noisy parallel data. Our intuition comes from the fact

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

[2] See for instance Equation 3 in Lample et al. (2018b).

that SMT generates less diverse translations, with a significantly different word frequency distribution than in translations naturally produced by humans. SMT is limited by the vocabulary of its phrase table and will favor the generation of frequent n-grams thanks to its language model. Same words appear more frequently in similar contexts, facilitating the training of word embeddings and compensating, to some extent, for the noisiness of the translations. In Appendix A, we provide results of our preliminary experiments supporting this assumption.

3 Experiments

> *Are BWE unsupervisedly and jointly trained on noisy synthetic data better than unsupervised mapped BWE?*

To answer this question, we conducted experiments in three different tasks with three language pairs: English–German (en-de), English–French (en-fr), and English–Indonesian (en-id).

3.1 Settings for training BWE

We trained monolingual word embeddings with `fastText` (Bojanowski et al., 2017)[3] separately on English (239M lines), German (237M lines), and French (38M lines) News Crawl corpora provided by WMT[4] for en-de and en-fr. For en-id, we used English (100M lines) and Indonesian (77M lines) Common Crawl corpora.[5] We then mapped the word embeddings into a BWE space using `VECMAP`,[6] one of the best and most robust methods for unsupervised mapping (Glavas et al., 2019). The resulting BWE were used as baselines in our evaluation tasks and also to bootstrap our USMT system.

Our initial USMT systems were induced with the following configuration. Maximum phrase length was set to six ($L = 6$). To make our experiments reasonably fast, we selected the 300k most frequent phrases referring to each monolingual corpus, and retained 300-best target phrases for each source phrase ($k = 300$). 4-gram language models were trained with `lmplz` (Heafield et al., 2013). Then, USMT systems were refined four times ($N = 4$) and used to generate synthetic parallel data by translating 10M sentences randomly sampled from the monolingual data. Finally, on the synthetic parallel data, we trained new BWE using `BIVEC`[7] with the parameters used in Upadhyay et al. (2016) and with word alignments determined by `fast_align` (Dyer et al., 2013).[8]

We performed contrastive experiments for some of our tasks with a simple method proposed by Levy et al. (2017), denoted `SENTID`,[9] with its default parameters for training BWE. `SENTID` does not optimize a joint objective but as for `BIVEC` we trained it on the synthetic parallel data and learned directly from scratch a single BWE space. `SENTID` does not require word alignments, but instead simply exploits sentence pair IDs as a bilingual signal associated with each word and train BWE by applying skipgram on a word/sentence-ID matrix.

All the methods for training word embeddings were trained with 512 dimensions and their `-min-count` parameter set to 5.

Note that in all our experiments, we filtered the vocabulary so that all BWE spaces have the same vocabulary when compared.

3.2 Task 1: Bilingual lexicon induction

Bilingual lexicon induction (BLI) is by far the most popular evaluation task for BWE used by previous work in spite of its limits (Glavas et al., 2019). In contrast to previous work, we used much larger test sets[10] for each language pair.

Table 1 reports on accuracy in retrieving a correct translation with CSLS (Lample et al., 2018a) for each source word of the test sets. For all the tasks, `BIVEC` and `SENTID` achieved better accuracy than `VECMAP`. This supports our assumption that even noisy synthetic parallel data can provide useful bilingual contexts for training BWE. The largest improvements were observed for en-id, with a gain of more than 10 points. Interestingly, `BIVEC` and `SENTID` performed similarly, pointing out that word alignments are not necessary in our scenario. The accuracy was higher when synthetic parallel data did not contain syn-

[3] https://github.com/facebookresearch/fastText
[4] http://www.statmt.org/wmt19/
[5] http://commoncrawl.org
[6] https://github.com/artetxem/vecmap
[7] https://github.com/lmthang/bivec
[8] https://github.com/clab/fast_align
[9] https://bitbucket.org/omerlevy/xling_embeddings
[10] https://github.com/facebookresearch/MUSE

Method	Data src-tgt	en→de	de→en	en→fr	fr→en	en→id	id→en
VECMAP	all-all	42.4	59.0	67.7	70.0	58.9	59.5
BIVEC	10M-0	**45.8**	59.2	73.9	71.3	**70.4**	69.7
SENTID	10M-0	**45.8**	60.1	**74.4**	71.8	69.8	69.2
BIVEC	0-10M	43.7	63.4	72.0	74.3	67.3	72.3
SENTID	0-10M	43.5	**63.5**	72.6	**74.8**	67.5	**73.4**
BIVEC	10M-10M	44.9	54.9	73.9	73.8	69.5	72.1
SENTID	10M-10M	45.4	62.1	74.2	74.0	69.4	73.0
Coverage ratio		15.1	14.7	24.8	26.9	27.8	25.4

Table 1: Accuracy in BLI for different BWE. The "Data" column indicates the number of sentences in the monolingual data used to train BWE: e.g., "0" means that the data of the corresponding language has been generated by USMT. For the last two rows, 20M synthetic sentence pairs have been used: 10M generated by L1→L2 and 10M generated by L2→L1 USMT systems. The last row indicate coverage ratio for each test set by the BWE. Best scores in each translation direction is presented in **bold**.

USMT	Data	en→de		en→fr		en→id	
	src-tgt	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU
Step 0	10M-0	47.1	(12.1)	74.1	(17.0)	65.2	(13.6)
Step 4	10M-0	46.4	(18.8)	75.6	(25.3)	69.4	(24.5)
Step 0	0-10M	43.8	(16.0)	72.8	(18.6)	64.6	(17.7)
Step 4	0-10M	44.0	(23.4)	73.5	(26.7)	66.4	(29.1)
Coverage ratio		14.3		23.1		23.0	

Table 2: Accuracy in BLI using BWE learned with BIVEC on synthetic parallel sentences generated either by step 0 or step 4 of USMT. BLEU scores of the USMT systems that generated the data were evaluated on the test sets presented in Section 3.3.

thetic English ("10M-0" for "en→∗" and "0-10M" for "∗→en"). Using the concatenation of the synthetic data generated by L1→L2 and L2→L1 (last two rows of the table) slightly underperformed the best configuration despite the use of twice more training data. This is presumably due to the presence of sentences of two very different natures, synthetic and original, in the same language.

To evaluate the robustness of BIVEC, we compared the performance to those obtained with noisier synthetic data generated by the initial USMT (without refinement). As shown in Table 2, we observed comparable results, especially for en→de and en→fr, confirming that this approach is very robust to noisy training data.

Although BIVEC and SENTID used a sub-part of the monolingual data used by VECMAP, their vocabulary size can be larger. This unintuitive observation comes from the use of USMT to generate synthetic data: L1 words not covered by the phrase table are directly copied in the translations. As a result, such L1 words are introduced into the L2 vocabulary even if they do not appear in the L2 monolingual data used to train VECMAP, artifically increasing the *coverage ratio*[11] of the lexicon. This side-effect of our method is especially useful for instance for named entities that should be kept as is. Since such words in L1 and their copies in L2 cooccur frequently in synthetic data, their embeddings are similar. Obviously, this side-effect is interesting only for close languages and may introduce numerous unwanted L1 words in the L2 space. See Appendix B for some more analyses.

3.3 Task 2: Machine translation

In the phrase table induction for USMT, both the geometry of the space (when retrieving the k-closest translations for a given source phrase) and the embeddings themselves (when computing cosine similarity for the translation probability) play an important role. Better BWE should lead to bet-

[11] As a definition for coverage, we chose the one implemented in VECMAP: the percentage of source words in a test bilingual lexicon that are in the vocabulary of the source word embeddings and that are paired with at least one target word that is in the vocabulary of the target word embeddings.

Method	en→de	en→fr	en→id
VECMAP	12.1	17.0	13.6
BIVEC	12.7	**17.3**	**15.9**
SENTID	**12.8**	**17.3**	15.8

Table 3: BLEU scores of USMT at step 0 with a phrase table induced using different BWE.

ter phrase tables and consequently translations of better quality. We thus regard USMT as an extrinsic evaluation task for BWE.

Table 3 shows BLEU scores for our USMT at step 0 on en-de Newstest2016, en-fr Newstest2014 of WMT, and en-id ALT (Riza et al., 2016)[12] test sets. We observed from 0.3 (BIVEC, en→fr) to 2.5 (BIVEC, en→id) BLEU points of improvements over USMT using VECMAP. Again, BIVEC and SENTID performed similarly. However, note that here USMT is merely an evaluation task: the improvement observed at step 0 are practically useless for USMT, since we can often gain much larger improvements through refinement as described in Section 2.2. Consequently, we assume that perfoming more iterations, i.e., retraining BWE on synthetic parallel data generated by an USMT system initialized from unsupervised joint BWE, will not improve either translation quality or BWE quality.

3.4 Task 3: Monolingual word analogy

In the literature, VECMAP and BIVEC BWE have been shown to perform as well as, or better than, word embeddings trained exclusively on monolingual data in monolingual tasks. Since we use significantly less and noisier data for training BIVEC than VECMAP, we assume that this observation may not hold in our configuration.

We tested our assumption with the English word analogy task of Mikolov et al. (2013b) by comparing VECMAP and BIVEC English word embeddings, with several different sets of en-fr synthetic parallel data for training BIVEC. As shown in Table 4, BIVEC led to significantly lower accuracy than VECMAP, especially for the configuration trained on synthetic English (generated from French) with a gap of 32.2 points. We also observed a lower accuracy when using original English, presumably due to the use of much smaller data than for training VECMAP. However,

[12] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/

Method	English data	Accuracy
VECMAP	239M (en)	77.8
BIVEC	10M (en→fr)	65.7
	10M (fr→en)	45.6
	10M (fr→en) + 10M (en→fr)	62.3
fastText	239M (en)	79.1
	10M (en→fr)	64.6
	10M (fr→en)	45.1
	10M (fr→en) + 10M (en→fr)	61.2

Table 4: Results on the English word analogy task using the English word embeddings.

when training monolingual word embeddings using fastText on the same English data used for training BIVEC, we observed that fastText underperforms BIVEC. This confirms that BIVEC can take advantage of noisy but bilingual contexts to monolingually improve word embeddings.

4 Conclusion and future work

We show in several cross-lingual NLP tasks that unsupervised joint BWE achieved better results than unsupervised mapped BWE. Our experiments also highlight the robustness of joint training that can take advantage of bilingual contexts even from very noisy synthetic parallel data. Since our approach works on top of unsupervised mapping for BWE and uses synthetic data generated by unsupervised MT, it will directly benefit from any future advances in these two types of techniques. Our approach has, however, a higher computational cost due to the need of generating synthetic parallel data, while generating more data would also improve the vocabulary coverage.

As a future work, we would like to study, for training BWE, the impact of the use of synthetic parallel data generated by unsupervised NMT, or of a different nature, such as translation pairs extracted from monolingual corpora without supervision. Such translation pairs are, in general, more fluent but potentially much less accurate.

Acknowledgments

We would like to thank the reviewers for their useful comments and suggestions. A part of this work was conducted under the program "Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology" of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

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

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *CoRR*, abs/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.

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 Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *CoRR*, abs/1902.00508.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jgou. 2018a. Word translation without parallel data. In *International Conference on Learning Representations*.

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

Omer Levy, Anders Søgaard, and Yoav Goldberg. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 765–774, Valencia, Spain. Association for Computational Linguistics.

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, Denver, Colorado. Association for Computational Linguistics.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian Language Treebank. In *Proceedings of the 2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)*, pages 1–6, Bali, Indonesia.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. 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)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1670, Berlin, Germany. Association for Computational Linguistics.

Data	en→de	en→fr
Europarl	52.5	75.9
Synthetic Europarl	50.4	74.7

Table 5: Accuracy of BWE jointly trained on the original and on the synthetic version of Europarl in bilingual lexicon induction tasks. Presented results are for the same vocabulary.

Method	Data		en→de		de→en		en→fr		fr→en		en→id		id→en	
	src	tgt	Cov.	Acc.	Cov.	Acc.	Cov.	Acc.	Cov.	Acc.	Cov.	Acc.	Cov.	Acc.
VECMAP	all	all	27.0	24.6	26.2	36.8	34.1	55.2	35.9	54.8	42.8	39.8	41.2	40.8
BIVEC	10M	0	24.3	60.6	22.9	70.0	32.4	74.3	33.1	73.5	35.6	73.3	32.6	74.6
SENTID	10M	0	24.3	60.5	22.9	70.5	32.4	75.0	33.1	73.8	35.6	72.9	32.6	74.3
BIVEC	0	10M	17.4	42.2	17.5	58.1	28.2	70.3	31.5	70.0	36.8	67.6	35.9	71.9
SENTID	0	10M	17.4	42.1	17.5	58.3	28.2	70.0	31.5	70.5	36.8	67.7	35.9	73.3
BIVEC	10M	10M	27.3	57.0	26.3	62.2	37.3	70.2	39.1	70.0	46.1	69.8	44.6	73.0
SENTID	10M	10M	27.3	57.3	26.3	66.7	37.3	70.8	39.1	70.5	46.1	70.1	44.6	74.6

Table 6: Results in BLI of VECMAP, BIVEC, and SENTID BWE, on the "full" Muse bilingual lexicons, without filtering the vocabulary. In other words, the compared BWE do not have the same vocabulary. The coverage is given by the VECMAP's evaluation script.

A Preliminary experiment

To empirically test our assumption on the robustness of BIVEC to noisiness of training data, we performed a preliminary experiment. First, we trained a low-quality SMT systems for en→de an d en→fr on small parallel corpora.[13] Then, a synthetic version of Europarl is compiled by coupling the English side of Europarl parallel corpora and its German and French translations generated by the SMT systems. Finally, with BIVEC, we obtained two types of BWE respectively from the original and the synthetic Europarl, and evaluated them in bilingual lexicon induction (BLI) tasks on the test sets used in Section 3.2.

Results are presented in Table 5. Despite the poor performance of our SMT systems, BWE learned from the synthetic Europarl were only slightly less accurate for BLI than the BWE learned from the original Europarl. This result supports our assumption that BIVEC can exploit noisy synthetic data produced by SMT.

B Bilingual lexicon induction: coverage statistics

To show how the vocabulary coverage varies between BWE spaces, and to evaluate their impact on the accuracy in BLI, we report in Table 6 the coverage and the accuracy in BLI for all the BWE evaluated without restricting their vocabulary to be the same. Note that, because of the differences in coverage, accuracy of joint BWE cannot directly be compared with VECMAP BWE.

[13] We used the News Commentary corpora provided by WMT for en→de and en→fr to train SMT systems performing at 15.4 and 20.1 BLEU points on Newstest2016 en-de and Newstest2014 en-fr, respectively.