Data augmentation for Bilingual Word Embeddings using Unsupervised Machine Translation

Anonymous COLING submission

Abstract

aaaa

1 Introduction

Bilingual word embeddings aim to learn a shared meaning space between two languages (the source and target languages) that can be useful for cross-lingual transfer learning (), or machine translation (Artetxe et al., 2018b; Lample et al., 2018a). Although early methods for bilingual word embeddings often utilize multilingual resources such as parallel corpora (Gouws et al., 2015; Luong et al., 2015) and word dictionaries (Mikolov et al., 2013), recent studies focus on completely unsupervised methods that do not require any cross-lingual supervision (Lample et al., 2018b; Artetxe et al., 2018a; Patra et al., 2019; Zhou et al., 2019). Most unsupervised methods fall into the category of the mapping-based method, which generally consists of the following procedures: train monolingual word embeddings independently in two languages; then, find a linear mapping that aligns the two embedding spaces. The mapping-based method is based on a strong assumption that the two independently trained embedding spaces have similar structures that can be aligned by a linear transformation, which has been, however, shown to be unlikely when the two corpora are from different domains or the two languages are typologically very different (Søgaard et al., 2018).

Current mapping-based methods only focus on how to align given two embeddings, and ignore how to improve the structural similarity of the source and target embeddings. In this paper, we propose such a method to facilitate cross-lingual embedding mapping, by augmenting the source or/and target corpora with the output from an unsupervised machine translation system (Artetxe et al., 2019), which do not require any parallel corpora.

The motivation of our method is three-fold. Firstly, the unsupervised mapping methods require comparable corpora, but the sufficient amount of comparable corpora is not readily available for every language pair. Our method can compensate the data amount to boost the performance of unsupervised bilingual word embeddings. Secondly, the source and target embeddings are usually trained on comparable corpora. The difference in the content of the two corpora may accentuate the structural difference between the two resulting embedding spaces, and thus we can mitigate that effect by making the source and target corpora parallel by automatically generated pseudo data. Thirdly, in the mapping-based method, the source and target embeddings are trained independently without taking into account the other language. Thus, the embedding structures may not be optimal for bilingual word embeddings. We argue that pseudo sentences generated by unsupervised machine translation system contain some trace of the original language, and using them when training mono-lingual embeddings can facilitate the structural correspondence of the two embeddings.

In the experiments using the Wikipedia Comparable corpora in English, French, German, and Japanese, we observe substantial improvements by our method in the task of bilingual lexicon induction without hurting the quality as monolingual embeddings. Moreover, we carefully analyze why our method improves the performance, and the result suggests that the generated translation data contains the information of the original language.

2 Background and Related Work

Bilingual Word Embeddings

Bilingual word embedding aims to learn semantic spaces shared between two languages. Mapping-based approaches are the method to transform ~~a~~ source word embeddings to the word embedding of the target language by using a linear projection with a transformation matrix W (Mikolov et al., 2013). This transformation matrix is learned using stochastic gradient descent by minimising mean squared error as described by the following equation. In this equation, x_i^s and x_i^t represent source/target seed word in the bilingual dictionary.

$$\omega_{MSE} = \sum_{i=1}^{n} \|Wx_i^s - x_i^t\|^2 \quad (1)$$

This method is based on the idea that word embeddings are structurally similar even if those of languages are different.

Recently, Artetxe et al (2018a) showed an unsupervised approach to learn bilingual word embeddings. In their method, training iterates through the following two steps:

1. induce an initial seed bilingual dictionary generated by using nearest neighbors based on the similarity between monolingual similarity distributions of words of each language

2. learn W to minimize the formula (1)

Most of unsupervised bilingual word embeddings methods are based on the mapping method, However, recent study shows that this assumption has been shown not to hold in general especially when the two corpora are from different domains or the two languages are typologically very different (Søgaard et al., 2018)

We assumed that the limitations could be due to two properties of the mapping-based method: (1) the source and target embeddings are independently learned without taking into account the other language. Thus, the embedding structures may not be optimal for bilingual word embeddings; (2) the source and target corpora are usually not parallel but comparable corpora. The difference in the content of the two corpora accentuate the structural difference between the two resulting embedding spaces.

To solve this problem, we propose a method to augment the source or/and target corpora with the output from an unsupervised machine translation model. We expect that the word co-occurrence information of each language when we learn the word embeddings will be similar by using the translated sentence as training data. It's because the generated data may contain information of other languages and the original sentence and the generated sentence can be considered as aligned parallel corpora, so the content become similar.

Unsupervised Machine Translation

Unsupervised machine translation is a machine translation model that can translate without any parallel corpus. First, word-by-word translation model is learned using bilingual word embedding. Then, language model which have monolingual grammar information are trained on both source and target languages using each comparable corpus. After that, initial machine translation $P^0_{s\to t}$ is created by word-by-word translation model and language model. Finally, $P^1_{t\to s}$ is learned in a supervised learning using the original training corpus and outputs of the corpus by $P^0_{s\to t}$ as a synthetic parallel corpora. In the same way, the quality of the translation model is improved with an iterative process. It can be said that improving the accuracy of the bilingual word embeddings improves the quality of the word-by-word translation, which leads to improvement of translation model.

As a unsupervised machine translation model, we chose unsupervised phrase-based SMT to generate pseudo corpus because it can generate better translations than unsupervised NMT especially on low-resource language. (Lample et al., 2018c) In the phrase-based SMT model, a phrase table is created for the word level translation as shown in the following equation.

$$p(t_j|s_i) = \frac{\exp\left(\frac{1}{T}\cos(e(t_j), e(s_i))\right)}{\sum_k \exp\left(\frac{1}{T}\cos(e(t_k), e(s_i))\right)} \quad (2)$$

describe hensu

Figure 1 shows overview of our simple method. We trained unsupervised machine translation, then input source/target train data to the machine translation and get pseudo corpus. Having done that, we concatenated pseudo corpus with the original training corpus, and learned bilingual word embedding again using the same procedure.

Related Work

Marie et al. (2019) proposed a ~~new~~ method to learn bilingual word embedding. First, they train unsupervised machine translation using train data, then generated synthetic parallel corpora. They then learned bilingual word embedding using word alignment and bilingual skip-gram (joint-learing).

We used the pseudo corpus not as a synthetic parallel corpus but as training data to learn monolingual word embedding of each language focusing on structure of word embeddings and mapping accuracy.

3 Experimental Design

We used the Wikipedia Comparable Corpora. All data except Japanese were lower-cased and tokenized with Mose's tokenizer. we tokenized japanese data with `kytea`[1] and nomerized with `mojimoji`. After that, we trained monolingual word embeddings with `fastText`[2] separately on four languages:English, French, German, Japanese. 512 dimensions and 5 -min-count parameter . 10M sentence We then mapped these word embeddings into shared embedding space using open source implementation `VecMap`[3].

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).

Using this BWE and language model, we initialized unsupervised phrase-based machine translation model. After that, we generated synthetic parallel data using all training data and improved the machine translation through iterative backtranslation. Table 10 shows BLEU scores of the unsupervised machine translation we use.

	en→fr	fr→en	en→de	de→en	en→ja	ja→en
3	**19.3**	**19.0**	**10.3**	**13.7**	3.6	1.4

Table 1: aaa

we input training corpus into the created unsupervised machine translation model, and obtained the output result (pseudo corpus). Having done that, we concatenated pseudo corpus with the original training corpus, and learned bilingual word embedding again using the same procedure.

In the joint-learning setting, we used the same training corpus and the unsupervised translation model as described above. We then genetate synthetic parallel data by all training corpus. On the synthetic parallel data, we learn word alignment by `fast_align` [4] (Dyer et al., 2013) with default parameters and trained bilingual word embeddings using `BIVEC` with the parameters used in (Upadhyay et al., 2016)

Note that in all our experiments, we filtered the vocabulary size of bilingual word embedding so that bilingual space have the same vocabulary when compared.

[1] `http://www.phontron.com/kytea/index-ja.html`
[2] `https://fasttext.cc`
[3] `https://github.com/artetxem/vecmap`
[4] `https://github.com/clab/fast_align`

method	source	target	en→fr	fr←en	en → de	de← en	en→ ja	ja←en
mapping	train	train	0.664	0.636	0.561	0.567	0.451	0.357
	train + pseudo	train	0.664	0.636	0.561	0.567	0.451	0.357
	train	train + pseudo	0.664	0.636	0.561	0.567	0.451	0.357
	train + pseudo	train + pseudo	0.664	0.636	0.561	0.567	0.451	0.357
joint-learning	train	pseudo	0.620	0.594	0.527	0.520	0.263	0.274
	pseudo	train	0.620	0.594	0.527	0.520	0.263	0.274
	train + pseudo	train + pseudo	0.620	0.594	0.527	0.520	0.263	0.274

Table 2: results of BLI

4 Experiments and results

4.1 Bilingual Lexicon Induction

Bilingual Lexicon Induction (BLI) is the standard task for evaluating bilingual word embeddings. Given a bilingual embedding space, the task is to retrieve the target word embedding from nearest neighbors of the source word embedding. In this experiment, we use mean reciprocal rank (MRR) (**?**)glavas-etal-2019-properly).

Table 2 shows that the BLI in the proposed method was improved compared to the existing method. The reason for improving accuracy may be that the translation reflects the nature of the source language. To verify this, we also experimented to augment the source language from a language other than the target language.

Corpus	fr	de	ja
en	0.655 ± 0.003	0.548 ± 0.004	0.451 ± 0.002
en + pseudo(fr)	$\mathbf{0.704} \pm 0.001$	0.561 ± 0.005	0.448 ± 0.002
en + pseudo(de)	0.647 ± 0.004	$\mathbf{0.610} \pm 0.004$	0.445 ± 0.005
en + pseudo(ja)	0.598 ± 0.002	0.496 ± 0.005	$\mathbf{0.463} \pm 0.002$

Table 3: Results of BLI

From the results, it was confirmed that the accuracy of the BLI was usually improved by performing the extension using the pseudo corpus, but it was confirmed that BLI accuracy may be reduced when the source training corpus is extended by a pseudo corpus generated from another language which is not target language. This result supports the hypothesis that the method is affected by the nature of the target language, not just data extension.

4.2 Eigen Similarity

We used eigen similarity proposed by (Søgaard et al., 2018) to verify the isomorphism of bilingual word embedding made by pseudo corpora. First, we normalize the embeddings and get the nearest neighbor graphs of the 3000 most frequent words in each language. We then calculate their Laplacian matrices L1 and L2 from those graphs and find the smallest k such that the sum of the k largest eigenvalues of each Laplacian matrices is $< 90\%$ of all eigenvalues. Finally, we calculate squared differences between the k largest eigenvalues λ L1 and L2 . Smaller eigen similarity means higher isomophism.

$$\delta = \sum_{i=1}^{k} (\lambda_{1i} - \lambda_{2i})^2 \tag{3}$$

Table shows results of Eigen Similarity. The results show that augmentation with the target language certainly improves isomorphism.

Corpus	fr	de	ja
en	0.655 ± 0.003	0.548 ± 0.004	0.451 ± 0.002
en + pseudo(fr)	$\mathbf{0.704} \pm 0.001$	0.561 ± 0.005	0.448 ± 0.002
en + pseudo(de)	0.647 ± 0.004	$\mathbf{0.610} \pm 0.004$	0.445 ± 0.005
en + pseudo(ja)	0.598 ± 0.002	0.496 ± 0.005	$\mathbf{0.463} \pm 0.002$

Table 4: Eigen Similarity

4.3 Downstream Tasks

document classification Recently, bilingual word embedding are almost exclusively evaluated on BLI task, but Glavaš et al. (2019) showed bilingual word embedding with high score in BLI do not always perform well in actual cross-lingual tasks. Therefore, we evaluate our embedding with cross-lingual document classification and sentiment analysis. In the downstream task, we implemented a convolutional network followed by a multilayer perceptrons classifier.

Document Classification

MLDoc

The task is a classification of the news article topics: CCAT (Corporate / Industrial), ECAT (Economics), GCAT (Government / Social), MCAT (Markets) .

Sentiment Analysis

Sentiment Analysis is a task of classifying a sentence as a positive opinion or a negative opinion in a review sentence. We used amazon review as a data set. This data consists of rating data from 1 to 5 for amazon products and review text. We defined rating values 1-2 as "negative" and 4-5 as "positive", and excluded 3.

Result and Discussion

Table shows the results of downstream tasks. It was confirmed that the bilingual word embedding by the proposed method could outperform existing methods in the document classification task, but the sentiment analysis task did not show a consistent tendency.

	en-fr	en-de	en-ja
train	$79.5 \pm 1.5\,(92.6)$	$79.0 \pm 1.5\,(91.7)$	$70.4 \pm 1.2\,(92.2)$
train + pseudo	$\mathbf{82.2}^{\dagger} \pm 1.5\,(93.3)$	$79.3 \pm 2.0\,(92.0)$	$\mathbf{71.6}^{\dagger} \pm 0.8\,(93.3)$

Table 5: Document Classification

	en-fr	en-de	en-ja
train	$69.1 \pm 1.1\,(71.8)$	$63.7 \pm 1.3\,(71.1)$	$63.5 \pm 1.1\,(70.7)$
train + pseudo	$69.5 \pm 0.8\,(71.9)$	$\mathbf{65.1}^{\dagger} \pm 0.8\,(70.2)$	$62.8 \pm 1.4\,(70.6)$

Table 6: Sentiment Analysis

4.4 Monolingual Word Similarity

In the proposed method, we used the output result of machine translation as a training data, but the translated sentence was not always accurate and there are some noises. Therefore, we tested the quality of word embedding with the Word Similarity task. This task evaluates the quality of monolingual word embedding by measuring the correlation of the cosine similarity between word pair similarity created manually and actual cosine similarity.

We used simverb-3500[5] consisting of 3500 verb pairs and men[6] consisted of 3000 frequent words extracted from web text.

corpus	simverb-3500	men
train	0.259 ± 0.006	0.763 ± 0.001
train + pseudo(fr)	0.260 ± 0.004	$\mathbf{0.767}^{\dagger} \pm 0.002$
train + pseudo(de)	0.253 ± 0.003	$\mathbf{0.768}^{\dagger} \pm 0.002$
train + pseudo(ja)	$0.220^{\dagger} \pm 0.001$	$0.760^{\dagger} \pm 0.002$

Table 7: results of Word Similarity

Table 7 shows the results of word similarity. It was shown that the quality of the monolingual word embedding in French and German, which are relatively linguistically similar to English, is maintained despite the presence of noise. On the contrary, it can be seen that the accuracy has decreased in Japanese linguistically far from English. In the proposed method, BLI is improved in spite of lowering the quality of monolingual in English and Japanese. From this, it is presumed that the word embeddings by this method has a structure specialized for bilingual word mapping.

4.5 Application to Machine Translation

Table 8 shows BLEU scores for unsupervised machine translation initialised with our bilingual word embedding at each iterative step. We observed that the BLEU score in the first step is high or but will not improve translation accuracy with more iterations. this results are also reported in

High diversity of the pseudo corpus makes it difficult to learn the translation model and it can make model robust to noise (Edunov et al., 2018). As described in section 3.3, The vocabulary of pseudo corpus is standardized to some extent. We compared the vocabulary size per word in the training corpus and the pseudo corpus used in this experiment (Table 9). The results showed that the pseudo corpus had less vocabulary per word than the training corpus. As a result, specific words are easily mapped in a bilingual word embedding using a pseudo corpus, and then the translation model makes it easier to translate phrases in more specific patterns.

	en→fr	fr→en	en→de	de→en	en→ja	ja→en
0		**14.7**	**10.3**	**13.7**	3.6	1.4
1	**16.7**	**18.8**	**10.3**	**13.7**	3.6	1.4
2	**18.8**	**19.2**	**10.3**	**13.7**	3.6	1.4
3	**19.2**	**19.1**	**10.3**	**13.7**	3.6	1.4

Table 8: aaa

4.6 data

5 Conclusion and Future Work

In this paper, we proposed a approach to learn bilingual word embeddings using pseudo corpus generated from unsupervised machine translation in a mapping method. Our bilingual word embeddings achieved better results than existing mapping method and joint-learning method in our low resource scenario. We showed

As a future work,

[5]http://people.ds.cam.ac.uk/dsg40/simverb.html
[6]https://staff.fnwi.uva.nl/e.bruni/MEN

	en-fr		en-de		en-ja	
	en	fr	en	de	en	ja
train	1.60×10^{-3}	1.63×10^{-3}	1.51×10^{-3}	3.78×10^{-3}	1.52×10^{-3}	1.03×10^{-3}
pseudo	0.57×10^{-3}	0.57×10^{-3}	0.66×10^{-3}	0.59×10^{-3}	0.19×10^{-3}	0.17×10^{-3}

Table 9: vocablary size per words

	en→fr	fr→en	en→de	de→en	en→ja	ja→en
train	**19.3**	**19.0**	**10.3**	**13.7**	3.6	1.4
pseudo	**19.3**	**19.0**	**10.3**	**13.7**	3.6	1.4
train + pseudo	**19.3**	**19.0**	**10.3**	**13.7**	3.6	1.4
train + train	**19.3**	**19.0**	**10.3**	**13.7**	3.6	1.4

Table 10: aaa

Acknowledgements

we have shown that:

-
-

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, July. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised Neural Machine Translation. In *Proceedings of the 5th International Conference on Learning Representations*, pages 1–12, February.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July. 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, June.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vuli. 2019. 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*, pages 710–721, Florence, Italy, July.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 748–756, Lille, France, July.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised Machine Translation Using Monolingual Corpora Only. In *Proceedings of the 5th International Conference on Learning Representations*, pages 1–14, February.

Guillaume Lample, Alexis Conneau, Marc Aurelio Ranzato, Ludovic Denoyer, and Herve Jegou. 2018b. Word Translation without Parallel Data. In *Proceedings of the International Conference on Learning Representations*.

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

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, June. Association for Computational Linguistics.

Benjamin Marie and Atsushi Fujita. 2019. Unsupervised joint training of bilingual word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3224–3230, Florence, Italy, July.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *Computing Research Repository*, arXiv:1309.4168.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. 2019. Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, July.

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, July. 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, August.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density matching for bilingual word embedding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June. Association for Computational Linguistics.