Data Augmentation with Unsupervised Machine Translation Improves the Structural Similarity of Cross-lingual Word Embeddings

Sosuke Nishikawa, Ryokan Ri and Yoshimasa Tsuruoka

The University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

`sosuke-nishikawa@nii.ac.jp`

`{li0123,tsuruoka}@logos.t.u-tokyo.ac.jp`

Abstract

Unsupervised cross-lingual word embedding (CLWE) methods learn a linear transformation matrix that maps two monolingual embedding spaces that are separately trained with monolingual corpora. This method relies on the assumption that the two embedding spaces are structurally similar, which does not necessarily hold true in general. In this paper, we argue that using a pseudo-parallel corpus generated by an unsupervised machine translation model facilitates the structural similarity of the two embedding spaces and improves the quality of CLWEs in the unsupervised mapping method. We show that our approach outperforms other alternative approaches given the same amount of data, and, through detailed analysis, we show that data augmentation with the pseudo data from unsupervised machine translation is especially effective for mapping-based CLWEs because (1) the pseudo data makes the source and target corpora (partially) parallel; (2) the pseudo data contains information on the original language that helps to learn similar embedding spaces between the source and target languages.

1 Introduction

Cross-lingual word embedding (CLWE) methods aim to learn a shared meaning space between two languages (the source and target languages), which is potentially useful for cross-lingual transfer learning or machine translation (Yuan et al., 2020; Artetxe et al., 2018b; Lample et al., 2018a). Although early methods for learning CLWEs 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 have focused on fully unsupervised methods that do not require any cross-lingual supervision (Lample et al., 2018b; Artetxe et al., 2018a; Patra et al., 2019). Most unsupervised methods fall into the

category of mapping-based methods, which generally consist 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 is unlikely to hold true when the two corpora are from different domains or the two languages are typologically very different (Søgaard et al., 2018). To address this problem, several studies have focused on improving the structural similarity of monolingual spaces before learning mapping (Zhang et al., 2019; Vulić et al., 2020), but few studies have focused on how to leverage the text data itself.

In this paper, we show that the pseudo sentences generated from an unsupervised machine translation (UMT) system (Lample et al., 2018c) facilitates the structural similarity without any additional cross-lingual resources. In the proposed method, the training data of the source and/or target language are augmented with the pseudo sentences (Figure 1).

We argue that this method facilitates the structural similarity between the source and target embeddings for the following two reasons. Firstly, the source and target embeddings are usually trained on monolingual 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. Secondly, 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 CLWEs. We argue that pseudo sentences generated by a UMT
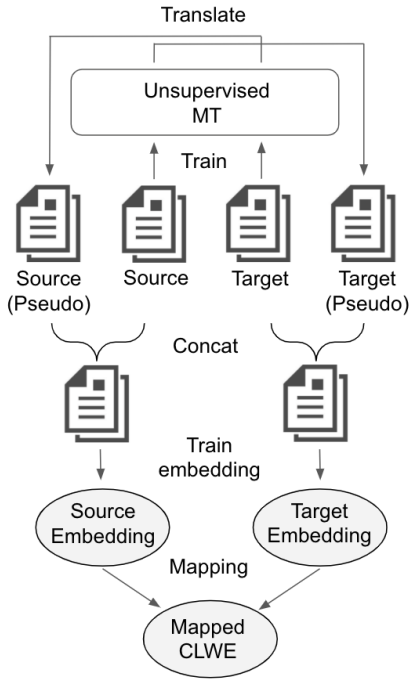
Figure 1: Our framework for training CLWEs using unsupervised machine translation (UMT). We first train UMT models using monolingual corpora for each language. We then translate all the training corpora and concatenate the outputs with the original corpora, and train monolingual word embeddings independently. Finally, we map these word embeddings on a shared embedding.

system contain some trace of the original language, and using them when training monolingual embeddings can facilitate the structural correspondence of the two sets of embeddings.

In the experiments using the Wikipedia dump in English, French, German, and Japanese, we observe substantial improvements by our method in the task of bilingual lexicon induction and downstream tasks without hurting the quality as monolingual embeddings. Moreover, we carefully analyze why our method improves the performance, and the result confirms that making the source and target corpora parallel does contribute to performance improvement, and also suggests that the generated translation data contain information about the original language.

2   Background and Related Work

Cross-lingual Word Embeddings

CLWE methods aim to learn a semantic space shared between two languages. Most of the current approaches fall into two types of methods: joint-training approaches and mapping-based approaches.

Joint-training approaches jointly train a shared embedding space given multilingual corpora with cross-lingual supervision such as parallel corpora (Gouws et al., 2015; Luong et al., 2015), document-aligned corpora (Vulic and Moens, 2016), or monolingual corpora along with a word dictionary (Duong et al., 2016).

On the other hand, mapping-based approaches utilize monolingual embeddings that are already obtained from monolingual corpora. They assume structural similarity between monolingual embeddings of different languages and attempt to obtain a shared embedding space by finding a transformation matrix $\mathbf{W}$ that maps source word embeddings to the target embedding space (Mikolov et al., 2013). The transformation matrix $\mathbf{W}$ is usually obtained by minimizing the sum of squared euclidian distances between the mapped source embeddings and target embeddings:

$$\underset{\mathbf{W}}{\operatorname{argmin}} \sum_{i}^{|D|} \|\mathbf{W}\mathbf{x}_i - \mathbf{y}_i\|^2, \qquad (1)$$

where D is a bilingual word dictionary that contains word pairs (x_i, y_i) and $\mathbf{x}_i$ and $\mathbf{y}_i$ represent the corresponding word embeddings.

Although finding the transformation matrix $\mathbf{W}$ is straightforward when a word dictionary is available, a recent trend is to reduce the amount of cross-lingual supervision or to find $\mathbf{W}$ in a completely unsupervised manner (Lample et al., 2018b; Artetxe et al., 2018a). The general framework of unsupervised mapping methods is based on heuristic initialization of a seed dictionary D and iterative refinement of the transformation matrix $\mathbf{W}$ and the dictionary D, as described in Algorithm 1. In our experiment, we use the unsupervised mapping-based method proposed by Artetxe et al. (2018a). Their method is characterized by the seed dictionary initialized with nearest neighbors based on similarity distributions of words in each language.

These mapping-based methods, however, are based on the strong assumption that the two independently trained embedding spaces have similar structures that can be aligned by a linear transformation. Although several studies have tackled improving the structural similarity of monolingual spaces before learning mapping (Zhang et al., 2019; Vulić et al., 2020), not much attention has been paid to how to leverage the text data itself.

Input: The source embeddings $\mathbf{X}$, the target embeddings $\mathbf{Y}$
Output: The transformation matrix $\mathbf{W}$
Heuristically induce an initial seed word dictionary D
while *not convergence* **do**
    Compute $\mathbf{W}$ given the word dictionary D from the equation (1)
    Update the word dictionary D by retrieving cross-lingual nearest neighbors in a shared
        embedding space obtained by $\mathbf{W}$
end
return $\mathbf{W}$

Algorithm 1: The general workflow of unsupervised mapping methods

In this paper, we argue that we can facilitate structural correspondence of two embedding spaces by augmenting the source or/and target corpora with the output from an unsupervised machine translation system (Lample et al., 2018c).

Unsupervised Machine Translation

Unsupervised machine translation (UMT) is the task of building a translation system without any parallel corpora (Artetxe et al., 2018b; Lample et al., 2018a,c; Artetxe et al., 2019b). UMT is accomplished by three components: (1) a word-by-word translation model learned using unsupervised CLWEs; (2) a language model trained on the source and target monolingual corpora; (3) a back-translation model where the model uses input and its own translated output as parallel sentences and learn how to translate them in both directions.

More specifically, the initial source-to-target translation model $P^0_{s \rightarrow t}$ is created by the word-by-word translation model and the language model of the target language. Then, $P^1_{t \rightarrow s}$ is learned in a supervised setting using the source original monolingual corpus paired with the synthetic parallel sentences of the target language generated by $P^0_{s \rightarrow t}$. Again, another source-to-target translation model $P^1_{s \rightarrow t}$ is trained with the target original monolingual corpus and the outputs of $P^0_{s \rightarrow t}$, and in the same way, the quality of the translation models is improved with an iterative process.

In our experiments, we adopt an unsupervised phrase-based statistical machine translation (SMT) method to generate a pseudo corpus because it produces better translations than unsupervised neural machine translation on low-resource languages (Lample et al., 2018c). The difference of the unsupervised SMT (USMT) model from its supervised counterpart is that the initial phrase table is derived based on the cosine similarity of unsupervised CLWEs, and the translation model is iteratively im-

proved by pseudo parallel corpora.

Our proposed method utilizes the output of a USMT system to augment the training corpus for CLWEs.

Exploiting UMT for Cross-lingual Applications

There is some previous work on how to use UMT to induce bilingual word dictionaries or improve CLWEs. Artetxe et al. (2019a) explored an effective way of utilizing a phrase table from a UMT system to induce bilingual dictionaries. Marie and Fujita (2019) generate a synthetic parallel corpus from a UMT system, and jointly train CLWEs along with the word alignment information (Luong et al., 2015). In our work, we use the synthetic parallel corpus generated from a UMT system not for joint-training but for data augmentation to train monolingual word embeddings for each language, which are subsequently aligned through unsupervised mapping. In the following sections, we empirically show that our approach leads to the creation of improved CLWEs and analyze why these results are achieved.

3 Experimental Design

In this section, we describe how we obtain mapping-based CLWEs using a pseudo parallel corpus generated from UMT. We first train UMT models using the source/target training corpora, and then translate them to the machine-translated corpora. Having done that, we simply concatenate the machine-translated corpus with the original training corpus, and learn monolingual word embeddings independently for each language. Finally, we map these embeddings to a shared CLWE space.

Corpora

We implement our method with two similar languages pairs: English-French (en-fr), English-German (en-de), and one distant language pair: English-Japanese (en-ja). We use plain texts from Wikipedia dumps[1], and randomly extract 10M sentences for each language. The English, French, and German texts are tokenized with the Moses tokenizer (Koehn et al., 2007) and lowercased. For Japanese texts, we use `kytea`[2] to tokenize and normalize them[3].

Training mapping-based CLWEs

Given tokenized texts, we train monolingual word embeddings using `fastText`[4] with 512 dimensions, a context window of size 5, and 5 negative examles. We then map these word embeddings on a shared embedding space using the open-source implementation `VecMap`[5] with the unsupervised mapping algorithm (Artetxe et al., 2018a).

Training UMT models

To implement UMT, we first build a phrase table by selecting the most frequent 300,000 source phrases and taking their 200 nearest-neighbors in the CLWE space following the setting of Lample et al. (2018c). We then train a 5-gram language model for each language with `KenLM` (Heafield et al., 2013) and combine it with the phrase table, which results in an unsupervised phrase-based SMT model. Then, we refine the UMT model through three iterative back-translation steps. At each step, we translate 100k sentences randomly sampled from the monolingual data set. We use a phrase table containing phrases up to a length of 4 except for initialization. The quality of our UMT models is indicated by the BLEU scores (Papineni et al., 2002) in Table 1. We use newstest2014 from WMT14[6] to evaluate En-Fr and En-De translation accuracy and the Tanaka corpus[7] for En-Ja evaluation.

---

[1] https://dumps.wikimedia.org/
[2] http://www.phontron.com/kytea/index-ja.html
[3] We convert all alphabets and numbers to half-width, and all katakana to full-width with the `mojimoji` library https://github.com/studio-ousia/mojimoji
[4] https://fasttext.cc
[5] https://github.com/artetxem/vecmap
[6] http://www.statmt.org/wmt14/translation-task.html
[7] http://www.edrdg.org/wiki/index.php/TanakaCorpus

en - fr		en - de		en - ja	
→	←	→	←	→	←
19.2	19.1	10.3	13.7	3.6	1.4

Table 1: BLEU scores of UMT.

Training CLWEs with pseudo corpora

We then translate all the training corpora with the UMT system and obtain machine-translated corpora, which we call *pseudo corpora*. We concatenate the pseudo corpora with the original corpora, and learn monolingual word embeddings for each language. Finally, we map these word embeddings to a shared CLWE space with the unsupervised mapping algorithm.

Models

We compare our method with a baseline with no data augmentation as well as the existing related methods: dictionary induction from a phrase table (Artetxe et al., 2019a) and the unsupervised joint-training method (Marie and Fujita, 2019). These two methods both exploit word alignments in the pseudo parallel corpus, and to obtain them we use `Fast_Align`[8] (Dyer et al., 2013) with the default hyperparameters. For the joint-training method, we adopt `bivec`[9] to train CLWEs with the parameters used in Upadhyay et al. (2016) using the pseudo parallel corpus and the word alignments. To ensure fair comparison, we implement all of these methods with the same UMT system.

4 Evaluation of Cross-lingual Mapping

In this section, we conduct a series of experiments to evaluate our method. We first evaluate the performance of cross-lingual mapping in our method (§ 4.1) and investigate the effect of UMT quality (§ 4.2). Then, we analyze why our method improves the bilingual lexicon induction (BLI) performance. Through carefully controlled experiments, we argue that it is not simply because of data augmentation but because: (1) the generated data makes the source and target corpora (partially) parallel (§ 4.3); (2) the generated data reflects the co-occurrence statistics of the original language (§ 4.4).

4.1 Bilingual Lexicon Induction

First, we evaluate the mapping accuracy of word embeddings using BLI. BLI is the task of iden-

---

[8] https://github.com/clab/fast_align
[9] https://github.com/lmthang/bivec

Method	source (en) orig.	source (en) psd.	target orig.	target psd.	en→fr MRR	en→fr P@1	fr→en MRR	fr→en P@1	en→de MRR	en→de P@1	de→en MRR	de→en P@1	en→ja MRR	en→ja P@1	ja→en MRR	ja→en P@1
BLI from phrase table	✓	-	-	✓	-	0.673	-	0.524	-	0.551	-	0.486	-	0.311	-	0.226
	-	✓	✓	-	-	0.509	-	0.697	-	0.302	-	0.542	-	0.198	-	0.259
	✓	✓	✓	✓	-	0.673	-	0.522	-	0.551	-	0.486	-	0.311	-	0.226
joint training	✓	-	-	✓	0.640	0.636	0.615	0.634	0.552	0.509	0.545	0.520	0.347	0.295	0.272	0.227
	-	✓	✓	-	0.587	0.579	0.643	0.685	0.535	0.491	0.577	0.549	0.279	0.226	0.305	0.249
	✓	✓	✓	✓	0.654	0.642	0.642	0.650	0.585	0.532	0.520	0.518	0.325	0.267	0.295	0.234
mapping	✓	-	✓	-	0.670	0.612	0.650	0.614	0.579	0.484	0.587	0.488	0.471	0.378	0.364	0.242
mapping (+ pseudo)	✓	-	✓	✓	0.709	0.666	0.687	0.688	**0.656**	**0.582**	0.635	**0.563**	**0.514**	**0.405**	**0.436**	**0.304**
	✓	✓	✓	-	**0.728**	**0.684**	**0.703**	**0.700**	0.647	0.566	0.636	0.562	0.486	0.392	0.407	0.297
	✓	✓	✓	✓	0.721	0.677	0.696	**0.700**	0.652	0.574	**0.637**	**0.563**	0.497	0.387	0.426	0.300

Table 2: Comparison with previous approaches in BLI. "orig." and "psd." indicate original training corpus and pseudo corpus. In each cell, the left cell shows the result of MRR, and the right cell shows the result of p@1.

tifying word translation pairs, and is a common benchmark for evaluating CLWE methods. In these experiments, we use Cross-Domain Similarity Local Scaling (Lample et al., 2018b) as the method for identifying translation pairs in the two embedding spaces. For BLI scores, we adopt the mean reciprocal rank (MRR) (Glavaš et al., 2019) and P@1.

We use XLing-Eval[10] as test sets for En-Fr and En-Ge. For En-Ja. We create the word dictionaries automatically using Google Translate[11], following Ri and Tsuruoka (2020). Other than BLI from a phrase table, we train three sets of embeddings with different random seeds and report the average of the results.

We compare the proposed method with other alternative approaches in BLI as shown in Table 2. In all the language pairs, the mapping method with pseudo data augmentation achieves better performance than the other methods. Here, one may think that the greater amount of data can lead to better performance, and thus augmenting both the source and target corpora shows the best performance. However, the result shows that it is not necessarily the case: for our mapping method, augmenting only either the source or target, not both, achieves the best performance in many language pairs. This is probably due to the presence of two pseudo corpora with different natures.

As for the two methods using word alignments (BLI from phrase table; joint training), we observe some cases where these models underperform the mapping methods, especially in English and Japanese pairs. We attribute this to our relatively low-resource setting where the quality of the synthetic parallel data is not sufficient to per-

BT step	en - fr BLI MRR	en - fr BLI P@1	en - fr BLEU	en - de BLI MRR	en - de BLI P@1	en - de BLEU
-	0.670	0.612	-	0.579	0.484	-
0	0.711	0.646	14.7	0.592	0.508	10.7
1	0.714	0.651	18.8	0.615	0.524	13.5
2	0.728	0.684	19.2	0.647	0.566	13.7

Table 3: Results of BLI score on CLWEs using pseudo corpus generated from different quality UMTs.

form these methods which require word alignment between parallel sentences.

4.2 Effect of UMT quality

To investigate the effect of UMT quality on our method, we compare the accuracy of BLI on the CLWEs using pseudo data generated from UMT models of different qualities. As a translator with low performance, we prepare models that perform fewer iterations on back-translation (BT). Note that we compare the results on the source-side (English) extension, where the quality of the translation is notably different. As shown in Table 3, we find that the better the quality of generated data, the better the performance of BLI.

4.3 Effect of sharing content

In the mapping method, word embeddings are independently trained by monolingual corpora that do not necessarily have the same content. As a result, the difference in the corpus contents can hurt the structural similarity of the two resulting embedding spaces. We hypothesize that using synthetic parallel data which have common contents for learning word embeddings leads to better structural correspondence, which improves cross-lingual mapping.

To verify the effect of sharing the contents using parallel data, we compare the extensions with a parallel corpus and a non-parallel corpus. More concretely, we first split the original training data

Extension		en - fr	en - de	en - ja
pseudo	parallel			
-	-	0.621 / 711	0.502 / 877	0.426 / 1776
×	×	0.630 / 838	0.509 / 1714	0.429 / 2301
✓	×	0.686 / **123**	0.569 / 272	0.454 / 1050
✓	✓	**0.695** / 144	**0.585** / 183	**0.459** / 1024

Table 4: Results of BLI score and eigenvector similarity. In each cell, the left cell shows the result of BLI, and the right cell shows the result of eigenvector similarity. Each row indicates, from top to bottom, no extension, extension with non-pseudo data, extension with non-parallel pseudo data, and extension with parallel pseudo data.

Corpus	fr-A	de-A	ja-A
en	0.621 / 711	0.502 / 877	0.426 / 1776
en + pseudo (fr-B)	**0.686 / 123**	0.516 / 315	0.421 / 2194
en + pseudo (de-B)	0.621 / 193	**0.569 / 272**	0.423 / 2173
en + pseudo (ja-B)	0.568 / 279	0.454 / 625	**0.454 / 1050**

Table 5: Results of BLI score and eigenvector similarity. Note that `lang`-A and pseudo (`lang`-B) are not parallel.

of the source and target languages evenly (each denoted as Split A and Split B). As the baseline, we train CLWEs with Split A. We use the translation of Split A of the target language data for the parallel extension of the source data, and Split B for the non-parallel extension. Also, we compare them with the extension with non-pseudo data, which is simply increasing the amount of the source language data by raw text.

Along with the BLI score, we show eigenvector similarity, a spectral metric to quantify the structural similarity of word embedding spaces (Søgaard et al., 2018). To compute eigenvector similarity, we normalize the embeddings and construct the nearest neighbor graphs of the 10,000 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 sum up the squared differences between the k largest eigenvalues from $L1$ and $L2$ and derive the eigen similarity. Note that smaller eigenvector similarity values mean higher degrees of structural similarity.

Table 4 shows the BLI scores and eigenvector similarity in each extension setting. The parallel extension method shows a slightly better BLI performance than the non-parallel extension. This supports our hypothesis that parallel pseudo data make word embeddings space more suitable for bilingual mapping because of sharing content. In eigenvector similarity, there is no significant improvement between the parallel and non-parallel corpora. This is probably due to large fluctuations in eigenvector similarity values. Surprisingly, the results show that augmentation using pseudo data

is found to be much more effective than the extension of the same amount of original training data. This result suggests that using pseudo data as training data is useful, especially for learning bilingual models.

4.4 Effect of reflecting the co-occurrence statistics of the language

We hypothesize that the translated sentences reflect the co-occurrence statistics of the original language, which makes the co-occurrence information on training data similar, improving the structural similarity of the two monolingual embeddings.

To verify this hypothesis, we experiment with augmenting the source language with sentences translated from a non-target language. To examine only the effect of the co-occurrence statistics of language and avoid the effects of sharing content, we use the extensions with the non-parallel corpus.

Table 5 shows that BLI performance and eigenvector similarity improve with the extension from the same target language, but that is not the case if the pseudo corpus is generated from a non-target language. These results indicate that our method can leverage learning signals on the other language in the pseudo data.

5 Downstream Tasks

Although CLWEs were evaluated almost exclusively on the BLI task in the past, Glavaš et al. (2019) recently showed that CLWEs that perform well on BLI do not always perform well in other cross-lingual tasks. Therefore, we evaluate our embeddings on the four downstream tasks: topic classification (TC), sentiment analysis (SA), dependency parsing (DP), and natural language inference

Task	en-fr			en-de			en-ja		
	mapping	mapping (+ pseudo)	joint training	mapping	mapping (+ pseudo)	joint training	mapping	mapping (+ pseudo)	joint training
TC	79.5 (92.6)	**82.2**[†] (93.3)	79.7 (92.5)	79.0 (91.7)	**79.3** (92.0)	70.4 (91.4)	70.4 (92.2)	**71.6**[†] (93.3)	66.7 (91.9)
SA	69.1 (71.8)	**69.5** (71.9)	66.3 (69.9)	63.7 (71.1)	**65.1**[†] (70.2)	62.5 (70.3)	**63.5** (70.7)	62.8 (70.6)	57.3 (66.8)
DP	63.9 (73.2)	**64.3** (73.5)	64.1 (75.1)	56.7 (73.2)	**57.0** (73.6)	55.9 (74.7)	17.8 (72.9)	**18.1** (73.3)	17.3 (74.8)
NLI	54.4 (70.3)	**54.7** (70.1)	45.0 (68.6)	55.7 (70.2)	**56.0** (70.3)	44.7 (69.7)	- -	- -	- -

Table 6: Results of Downstream tasks. Numbers in parentheses indicate the score of English validation data. The scores indicate averages of 20 experiments with different seeds. Statistically significant correlations are marked with a dagger (p <0.01).

(NLI).

Topic Classification This task is classifying the topics of news articles. We use the MLDoc [12] corpus compiled by Schwenk and Li (2018). It includes four topics: CCAT (Corporate / Industrial), ECAT (Economics), GCAT (Government / Social), MCAT (Markets). As the classifier, we implemented a simple light-weight convolutional neural network (CNN)-based classifier.

Sentiment Analysis In this task, a model is used to classify sentences as either having a positive or negative opinion. We use the Webis-CLS-10 corpus [13]. This data consists of review texts for amazon products and their ratings from 1 to 5. We cast the problem as binary classification and define rating values 1-2 as "negative" and 4-5 as "positive", and exclude the rating 3. Again, we use the CNN-based classifier for this task.

Dependency Parsing We train the deep biaffine parser (Dozat and Manning, 2017) with the UD English EWT dataset[14] (Silveira et al., 2014). We use the PUD treebanks[15] as test data.

Natural Language Inference We use the English MultiNLI corpus (Williams et al., 2018) for training and the multilingual XNLI corpus for evaluation (Conneau et al., 2018). XNLI only covers French and German from our experiment. We train the LSTM-based classifier (Bowman et al., 2015), which encodes two sentences, concatenated the representations, and then feed them to a multi-layer perceptron.

In each task, we train the model using English training data with the embedding parameters fixed . We then evaluate the model on the test data in other target languages.

Result and Discussion

Table 6 shows the test set accuracy of downstream tasks. For topic classification, our method obtains the best results in all language pairs. Especially in En-Fr and En-Ja, a significant difference is obtained in Student's t-test. For sentiment analysis, we observe a significant improvement in En-De, but cannot observe consistent trends in other languages. For dependency parsing and natural language inference, we observe a similar trend where the performance of our method outperforms other methods, although no significant difference is observed in the t-test. The cause of the lower performance of joint-training compared with the mapping method is presumably due to the poor quality of synthetic parallel data as described in § 4.1. In summary, given the same amount of data, the CLWEs obtained from our method tend to show higher performance not only in BLI but also in downstream tasks compared with other alternative methods, although there is some variation.

6 Analysis

Monolingual Word Similarity Our method uses a noisy pseudo corpus to learn monolingual word embeddings, and it might hurt the quality of monolingual embeddings. To investigate this point, we evaluate monolingual embeddings with the word similarity task. This task evaluates the quality of monolingual word embeddings by measuring the correlation between the cosine similarity in a vector space and manually created word pair similarity. We use simverb-3500[16] (Gerz et al.,

---

[12]https://github.com/facebookresearch/ MLDoc
[13]https://webis.de/data/webis-cls-10. html
[14]https://universaldependencies.org/treebanks/en_ewt/index.html
[15]https://universaldependencies.org/conll17/

[16]http://people.ds.cam.ac.uk/dsg40/simverb.html

	en-fr		en-de		en-ja	
corpus	en	fr	en	de	en	ja
origin	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 7: Type-token ratio of the training corpus (origin) and the pseudo-corpus (pseudo)

corpus	simverb-3500	men
en	0.259	0.763
en + pseudo (fr)	0.260	0.767
en + pseudo (de)	0.253	0.768
en + pseudo (ja)	0.220	0.760

Table 8: Results of word similarity. The scores indicate averages of 3 experiments with different seeds.

BT step	en→fr CLWE (no pseudo)	fr→en CLWE (no pseudo)	en→fr CLWE (+ pseudo)	fr→en CLWE (+ pseudo)
0		14.7		**14.8**
1	**16.7**	**18.8**	16.1	18.2
2	**18.8**	**19.2**	18.2	18.5
3	**19.2**	**19.1**	18.6	18.8

Table 9: BLEU scores of UMT at each back-translation step in En-Fr with a phrase table induced using different CLWEs.

2016) consisting of 3500 verb pairs and men[17] (Bruni et al., 2014) consisting of 3000 frequent words extracted from web text.

Table 8 shows the results of word similarity. The scores of monolingual word embeddings using a French and German pseudo corpus are maintained or improved, while they decrease in Japanese. This suggests that the quality of monolingual word embeddings could be hurt due to the low quality of the pseudo corpus or differences in linguistic nature. Nevertheless, the proposed method improves the performance of En-Ja's CLWE, which suggests that the monolingual word embeddings created with a pseudo corpus have a structure optimized for cross-lingual mapping.

Application to UMT UMT is one of the important applications of CLWEs. Appropriate initialization with CLWEs is crucial to the success of UMT (Lample et al., 2018c). To investigate how CLWEs obtained from our method affect the performance of UMTs, we compare the BLEU scores of UMTs initialized with CLWEs with and without a pseudo corpus at each iterative step. As shown in Table 9, we observe that initialization with CLWE using the pseudo data result in a higher BLEU score in the first step but does not improve the score at further steps compared to the CLWE without the pseudo data. Marie and Fujita (2019) also demonstrate the same tendency in the CLWE with joint-training.

To investigate this point, we compare the lexical densities of the training corpus and the pseudo-corpus used in the above experiments (§ 4, 5) using type-token ratio (Table 7). The results demonstrate that the pseudo corpus has a smaller vocabulary per word than the training corpus, and thus it is

standardized to some extent as reported in Vanmassenhove et al. (2019). As a result, specific words might be easily mapped in CLWEs using a pseudo corpus[18], and then the translation model makes it easier to translate phrases in more specific patterns. Hence, the model cannot generate diverse data during back-translation, and the accuracy is not improved due to easy learning.

7 Conclusion and Future Work

In this paper, we show that training cross-lingual word embeddings with pseudo data augmentation improves performance in BLI and downstream tasks. We analyze the reason for this improvement and found that the pseudo corpus reflects the co-occurrence statistics and content of the other language and that the property makes the structure of the embedding suitable for cross-lingual word mapping.

Recently, Vulić et al. (2019) have shown that fully unsupervised CLWE methods fails in many language pairs and argue that researchers should not focus too much on the fully unsupervised settings. Still, our findings that improve structural similarity of word embeddings in the fully unsupervised setting could be useful in semi-supervised settings, and thus we would like to investigate this direction in the future.

---

[17]https://staff.fnwi.uva.nl/e.bruni/MEN

[18]In a preliminary experiment, we investigated the variation in performance of cross-lingual mapping with and without pseudo according to the frequency of words in the source language, but there was little correlation between them.

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019a. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007.

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

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

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Timothy Dozat and Christopher D Manning. 2017. Deep Biafine Attention for Neural Dependency Parsing. In *Proceedings of the International Conference on Learning Representations*.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning Crosslingual Word Embeddings without Bilingual Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295.

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.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721.

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.

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 (Volume 2: Short Papers)*, pages 690–696.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej 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.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

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

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.

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.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193.

Ryokan Ri and Yoshimasa Tsuruoka. 2020. Revisiting the context window for cross-lingual word embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 995–1005.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904.

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.

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.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418.

Ivan Vulić, Anna Korhonen, and Goran Glavaš. 2020. Improving bilingual lexicon induction with unsupervised post-processing of monolingual word vector spaces. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 45–54.

Ivan Vulic and Marie-Francine Moens. 2016. Bilingual Distributed Word Representations from Document-aligned Comparable Data. *Journal of Artificial Intelligence Research*, 55(1):953–994.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive refinement of cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5984–5996.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3180–3189.

8 Appendix

A The hyperparameters for downstream tasks

A.1 Document Classification and Sentiment Analysis

hyperparameters		
CNN Classifier	number of filters	8
	ngram_filter_sizes	2, 3, 4, 5
	MLP hidden size	32
Training	optimizer	Adam
	learning rate	0.001
	lr scheduler	halved each time the dev score stops improving
	patience	3
	batch size	50

A.2 Dependency Parsing

hyperparameters		
Graph-based Parser	LSTM hidden size	200
	LSTM number of layers	3
	tag representation dim	100
	arc representation dim	500
	pos tag embedding dim	50
Training	optimizer	Adam
	learning rate	0.001
	lr scheduler	halved each time the dev score stops improving
	patience	3
	batch size	32

A.3 Natural Language Inference

hyperparameters		
Sentence Encoder	LSTM hidden size	300
	LSTM number of layers	2
Training	optimizer	Adam
	learning rate	0.001
	lr scheduler	halved each time the dev score stops improving
	patience	3
	batch size	64