Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models

Takashi Wada[1], **Tomoharu Iwata**[2,3], and **Yuji Matsumoto**[1,3]

[1]Nara Institute of Science and Technology
[2]NTT Communication Science Laboratories
[3]RIKEN Center for Advanced Intelligence Project (AIP)
[1]{wada.takashi.wp7,matsu}@is.naist.jp
[2]tomoharu.iwata.gy@hco.ntt.co.jp

Abstract

Recently, a variety of unsupervised methods have been proposed that map pre-trained word embeddings of different languages into the same space without any parallel data. These methods aim to find a linear transformation based on the assumption that monolingual word embeddings are approximately isomorphic between languages. However, it has been demonstrated that this assumption holds true only on specific conditions, and with limited resources, the performance of these methods decreases drastically. To overcome this problem, we propose a new unsupervised multilingual embedding method that does not rely on such assumption and performs well under resource-poor scenarios, namely when only a small amount of monolingual data (i.e., 50k sentences) are available, or when the domains of monolingual data are different across languages. Our proposed model, which we call 'Multilingual Neural Language Models', shares some of the network parameters among multiple languages, and encodes sentences of multiple languages into the same space. The model jointly learns word embeddings of different languages in the same space, and generates multilingual embeddings without any parallel data or pre-training. Our experiments on word alignment tasks have demonstrated that, on the low-resource condition, our model substantially outperforms existing unsupervised and even supervised methods trained with 500 bilingual pairs of words. Our model also outperforms unsupervised methods given different-domain corpora across languages. Our code is publicly available[1].

1 Introduction

Learning cross-lingual or multilingual word embedding has been recognised as a very important research topic in natural language processing (NLP). Its objective is to map monolingual word embeddings of different languages into a common space, and this research has been applied to many multilingual tasks such as machine translation (Zou et al., 2013) and bilingual named entity recognition (Rudramurthy et al., 2016). It also enables the transfer of knowledge from one language into another (Xiao and Guo, 2014; Adams et al., 2017).

A number of supervised and unsupervised methods have been proposed that obtain cross-lingual word embeddings. Both supervised and unsupervised methods aim to find such a linear transformation that maps word embeddings in a source language into a target language space. Supervised methods employ bilingual dictionaries to learn the mapping (Mikolov et al., 2013b; Xing et al., 2015; Smith et al., 2017; Artetxe et al., 2018a), while unsupervised ones utilise the similarities or distance of word embeddings spaces across different languages (Conneau et al., 2018; Zhang et al., 2017a; Xu et al., 2018; Artetxe et al., 2018b).

Since the common objective of most of the supervised and unsupervised methods is to find an orthogonal linear mapping between languages, they heavily rely on the assumption that monolingual word embeddings are approximately isomorphic. However, Søgaard et al. (2018) have found that this assumption does not hold true in general, and demonstrated that it requires three specific conditions for the unsupervised method of Conneau et al. (2018) to perform well. The conditions are: Languages to align are linguistically similar; Monolingual word embeddings are trained by the same algorithms; And the domains of the monolingual corpora are similar across languages. In particular, the last condition is hard to assume when dealing with resource-poor languages, for which unsupervised methods can be

[1] https://github.com/twadada/multilingual-nlm

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3113–3124
Florence, Italy, July 28 - August 2, 2019.

beneficial in reality.

To overcome the limitations of previous work, we propose a new unsupervised multilingual word embedding method called *Multilingual Neural Language Model* (MNLM). In what follows, we summarise our main contributions and novelty of our proposed model:

Contributions

- We have discovered another limitation of the existing unsupervised methods: They do not perform well under the low-resource condition, namely when only small monolingual corpora (i.e., 50k sentences) are available in source and/or target languages. We have also confirmed that word embeddings are far from being isomorphic across languages under this condition, indicating that the conventional approachs are effective only for resource-rich languages. This is a serious problem since unsupervised learning is supposed to be beneficial when dealing with low-resource languages.

- We propose a new unsupervised multilingual word embedding method that overcomes the limitations of the existing methods. Our approach can successfully obtain multilingual word embeddings under the challenging conditions when only small monolingual corpora are available, or when the domains of the monolingual corpora are different across languages (we define these conditions as 'low-resource condition' and 'different-domain condition', respectively).

Novelty of Our Proposed Model

Whereas the existing unsupervised methods aim to map pre-trained word embeddings between languages based on the strong assumption that monolingual word embeddings are approximately isomorphic, our method does not require such assumption or pre-trained word embeddings; instead, it learns multilingual word embeddings jointly using forward and backward LSTM language models (Mikolov et al., 2010). Our model shares the language models among multiple languages and aims to learn a common sequential structure of different languages such as a common basic word order rule (e.g., subject-verb-object). The word embeddings of each language are trained independently, but sharing the LSTM networks encourages the embeddings to be mapped into the same space, generating multilingual word embeddings. Our experiments show that our unique approach makes it possible to obtain multilingual word embeddings with limited resources.

2 Related Work

Mikolov et al. (2013b) have proposed to obtain cross-lingual word representations by learning a linear mapping between two monolingual word embedding spaces. Later, Xing et al. (2015) have shown that enforcing an orthogonality constraint on the mapping improves the performance, and that offers a closed form Procrustes solution obtained from the singular value decomposition (SVD) of YX^T

$$\begin{aligned} W^* &= \arg\min_{W} \|WX - Y\|^2 = UV^{\mathrm{T}}, \\ s.t.\ \ & U\Sigma V^{\mathrm{T}} = \mathrm{SVD}(YX^{\mathrm{T}}), \end{aligned} \quad (1)$$

where W is a mapping matrix and Σ is a diagonal matrix.

Following this work, a variety of unsupervised methods have been proposed that obtain cross-lingual representations without any bilingual supervision. Zhang et al. (2017a) have proposed an unsupervised method that obtains the linear transformation using adversarial training (Goodfellow et al., 2014): during the training, a discriminator is trained to distinguish between the mapped source embeddings and the target embeddings, while the mapping matrix is trained to fool the discriminator. Conneau et al. (2018) employ a similar approach to Zhang et al. (2017a); they acquire an initial matrix using adversarial training and refine it by solving the orthogonal Procrustes problem. Zhang et al. (2017b) and Xu et al. (2018) obtain cross-lingual representations by minimising the earth-mover's distance and Sinkhorn distance, respectively. Artetxe et al. (2018b) propose an unsupervised self-learning method. Their method starts from roughly aligning words across languages using structural similarities of word embedding spaces, and refines the word alignment by repeating a robust self-learning method until convergence. They show that their approach is more effective than Zhang et al. (2017a) and Conneau et al. (2018) when languages to align are distant or monolingual corpora are not comparable across language. Recently, Chen and Cardie (2018) and Alaux et al. (2018) have proposed unsupervised multilingual word embedding methods. Their methods map word embeddings of more than two languages into a common space by capturing the inter-dependencies among multiple languages.

3 Our Model

3.1 Overview

We propose a new unsupervised multilingual word embeddings method called Multilingual Neural Language Model. Fig.1 briefly illustrates our proposed model. The model consists of bidirectional language models similar to ELMo (Peters et al., 2018), and most of the parameters are shared among multiple languages. In what follows, we summaries which parameters are shared across languages or specific to each language:

- Shared Parameters
  - $\overrightarrow{f}$ and $\overleftarrow{f}$: LSTM networks which perform as forward and backward language models, independently.
  - $\boldsymbol{E}^{\mathbf{BOS}_{\mathbf{fwd}}}$ and $\boldsymbol{E}^{\mathbf{BOS}_{\mathbf{bkw}}}$: The embeddings of initial inputs to the forward and backward language models, respectively.
  - $\boldsymbol{W}^{\mathbf{EOS}}$: The linear mapping for <EOS>, which is used to calculate the probability of the end of a sentence at every time-step.
- Specific Parameters to Language ℓ
  - $\boldsymbol{E}^{\boldsymbol{\ell}}$: Word embeddings of language ℓ
  - $\boldsymbol{W}^{\boldsymbol{\ell}}$: Linear projection of language ℓ, which is used to calculate the probability distribution of the next word.

The LSTMs $\overrightarrow{f}$ and $\overleftarrow{f}$ are shared among multiple languages and capture a common language structure. On the other hand, the word embeddings E^{ℓ} and linear projection W^{ℓ} are specific to each language ℓ. Since different languages are encoded by the same LSTM functions, similar words across different languages should have a similar representation so that the shared LSTMs can encode them effectively. For instance, suppose our model encodes an English sentence 'He drives a car.' and its Spanish translation 'El conduce un coche.' In these sentences, each English word corresponds to each Spanish one in the same order. Therefore, these equivalent words would have similar representations so that the shared language models can encode the English and Spanish sentences effectively. Although in general, each language has its different grammar rules, the shared language models are trained to roughly capture the common structure such as common basic word order rules (e.g., subject-verb-object) among different languages. Sharing <BOS> and <EOS> symbols ensures that the beginning and end of the hidden states are in the same space regardless of language, which encourages the model to obtain multilingual representations.

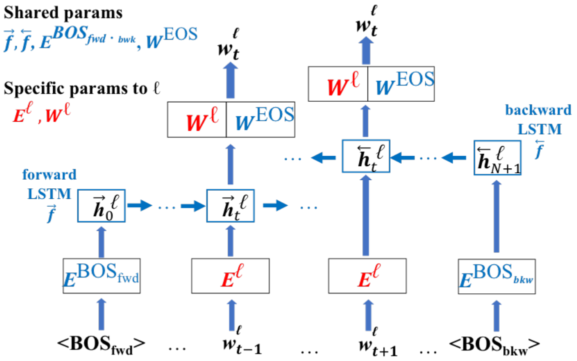

Figure 1: Illustration of our proposed model *Multilingual Neural Language Models*.

The limitation of our model is that it is only applicable to the languages that have common word order rules such as subject-verb-object and subject-object-verb. Although this limitation may sound somewhat significant, our experiments show that our model performs well not only for closely related language pairs such as French-English but also for linguistically distant languages such as English-Finnish[2] and Turkish-Japanese. In fact, our experiments show that it is extremely difficult for the existing unsupervised methods as well as for our model to align very distant languages which have different word order, such as English and Japanese.

3.2 Network Structure

Suppose a sentence with N words in language ℓ, $\langle w_1^{\ell}..., w_N^{\ell}\rangle$. The forward and backward language models calculate the probability of a next word w_t^{ℓ}

[2] Finnish is often considered as a non-Indo-European synthetic language, whereas English is often regarded as an Indo-European analytic language.

given the previous words:

$$p(w^\ell_1 ..., w^\ell_N) = \prod_{t=1}^{N} p(w^\ell_t | w^\ell_1 ..., w^\ell_{t-1}). \qquad (2)$$

$$p(w^\ell_1 ..., w^\ell_N) = \prod_{t=1}^{N} p(w^\ell_t | w^\ell_{t+1} ..., w^\ell_N). \qquad (3)$$

The tth hidden states h^ℓ_t of the forward and backward language models are calculated based on the previous hidden state and word embedding,

$$\overrightarrow{h}^\ell_t = \overrightarrow{f}(\overrightarrow{h}^\ell_{t-1}, x^\ell_{t-1}), \qquad (4)$$

$$\overleftarrow{h}^\ell_t = \overleftarrow{f}(\overleftarrow{h}^\ell_{t+1}, x^\ell_{t+1}), \qquad (5)$$

$$x^\ell_t = \begin{cases} E^{\mathrm{BOS}_{\mathrm{fwd}}} & \text{if } t = 0\ , \\ E^{\mathrm{BOS}_{\mathrm{bkw}}} & \text{if } t = N{+}1, \\ E^\ell(w^\ell_t) & \text{otherwise}, \end{cases} \qquad (6)$$

where $\overrightarrow{f}(\cdot)$ and $\overleftarrow{f}(\cdot)$ are the standard LSTM functions. Note that the same word embedding function E^ℓ is used among the forward and backward language models. The probability distribution of the upcoming word w^ℓ_t is calculated by the forward and backward models independently based on their current hidden state:

$$p(w^\ell_t | w^\ell_1 ..., w^\ell_{t-1}) = \mathrm{softmax}(g^\ell(\overrightarrow{h}^\ell_t))), \qquad (7)$$

$$p(w^\ell_t | w^\ell_{t+1} ..., w^\ell_N) = \mathrm{softmax}(g^\ell(\overleftarrow{h}^\ell_t)), \qquad (8)$$

$$g^\ell(h^\ell_t) = [W^\ell(h^\ell_t), W^{\mathrm{EOS}}(h^\ell_t)], \qquad (9)$$

where $[x, y]$ means the concatenation of x and y. W^{EOS} and W^ℓ are matrices with the size of $(1 \times d)$ and $(V^\ell \times d)$, where d is the size of hidden state and V^ℓ is the vocabulary size of language ℓ excluding <EOS>. As with the word embeddings, those matrices are shared among the forward and backward language models.

The proposed model is trained by maximising the log likelihood of the forward and backward directions for each language ℓ:

$$\sum_{l=1}^{L} \sum_{i=1}^{S^\ell} \sum_{t=1}^{N^i} \log p(w^\ell_{i,t} | w^\ell_{i,1} ... w^\ell_{i,t-1}; \overrightarrow{\theta}) + \log p(w^\ell_{i,t} | w^\ell_{i,t+1} ... w^\ell_{i,N^i}; \overleftarrow{\theta}),$$

where L and S^ℓ denote the numbers of languages and sentences of language ℓ. $\overrightarrow{\theta}$ and $\overleftarrow{\theta}$ denote the parameters for the forward and backward LSTMs $\overrightarrow{f}$ and $\overleftarrow{f}$, respectively.

4 Experiments

4.1 Data and Experimental Conditions

We trained our model and baselines under the following two conditions:

1. **low-resource condition**: Only small monolingual corpora are available.
2. **different-domain condition**: Relatively large monolingual corpora are available but their domains are different across languages.

On each condition, we conducted cross-lingual and multilingual embedding experiments, respectively.

4.1.1 Cross-lingual Word Embedding

In the experiments of cross-lingual embedding, we evaluated the quality of cross-lingual embeddings between seven pairs of source-target languages: {German, Spanish, French, Russian, Czech, Finnish, Japanese}-English. For the low-resource condition, we used subsets of News Crawl monolingual corpora[3]. We used 50k sentences for source languages, and either 50k or 1M sentences for the target language (i.e., English). This condition simulates two realistic scenarios; the case when analysing inter-dependencies between multiple minor languages, or between minor and major languages.

For the different-domain condition, we added {Tamil, Turkish}-Japanese pairs and North Saami-{Finnish, English} pairs to the seven pairs described above. North Saami is one of the minor languages spoken in northern Finland, Sweden and Norway, and it is so close to Finnish that transfer learning between them is very effective in dependency parsing (Lim et al., 2018). Note that the basic word order of Tamil, Turkish and Japanese is subject-object-verb (SOV), while the one of the other languages is SVO. We used Europarl corpus (Koehn, 2005) for English, Wikipedia for Japanese, SIKOR North Saami corpus[4] for North Saami, and news data for the other languages[5]. We extracted 1M sentences

[3] downloaded from http://www.statmt.org and http://wortschatz.uni-leipzig.de/en/download

[4] https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/8AK7KZ

[5] The vocabulary sizes of Europarl and News Crawl corpora in English are significantly different (79,258 v.s. 265,368 words), indicating the major differences between these domains

from these corpora except for North Saami, for which we used the whole corpus which contains 0.75M sentences. This different-domain condition also simulates the cases of analysing inter-dependencies among minor languages; large monolingual data containing up to 1M sentences may be available in each language, but it is hard to assume that their domains are similar across languages.

4.1.2 Multilingual Word Embedding

We trained multilingual word embeddings among the four linguistically similar languages: German, Spanish, French, and English. We conducted experiments under the following three conditions: (a) 50k sentences in News Crawl are used for each language; (b) 50k sentences in News Crawl are used for German, Spanish, French and 1M for English; (c) 1M sentences in News Crawl are used for German, Spanish, French and 1M sentences in Europarl for English.

4.2 Evaluation

In our experiment, we evaluated cross-lingual and multilingual word embeddings on the word alignment tasks. In the cross-lingual experiments, we used 1000 unique pairs of words in the dictionaries and we report p@5 in each language. That is, for each word in the 1000 source words, we extracted the 5 most similar words from the 1000 target words and checked how often the correct translation is included in them. In the multilingual experiments, we extracted 500 words aligned among English, French, Spanish and German and evaluated p@5 of 'joint' alignment among the four languages. That is, for each English word we extracted the 5 most similar words in French, German and Spanish independently, and evaluated how often the correct translation of the English word is included in all of the three languages. In most language pairs, these 1000 and 500 words were extracted from bilingual dictionaries published by Conneau et al. (2018) so that they did not contain any unknown words in all the training settings[6]. For North Saami-{Finnish, English}, we used the North Saami-Finnish dictionary[7] used by Lim et al. (2018) and aligned it with a Finnish-English dictionary published by Conneau et al. (2018) to build a North Saami-English dictionary.

When only 50k sentences were used both for the source and target languages, we trained all the models three times with different random seeds and calculated the average precision in both the cross-lingual and multilingual experiments. This is because unsupervised learning with small data can be unstable.

4.3 Baseline

Baseline models aim to map pre-trained word embeddings of different languages into a common space. For a fair comparison to our model, we used word2vec (Mikolov et al., 2013a), that pre-train word embeddings at a token level. We used their code with the default setting[8] except for the embedding size and minimum frequency, which were set the same as our model. Note that these pre-trained embeddings were used only by baseline models, not by ours.

As baselines of cross-lingual word embedding methods, we chose Xu et al. (2018), Artetxe et al. (2018b), and Conneau et al. (2018) with and without normalisation. We also compared our model against (weakly) supervised cross-lingual word embedding methods (Artetxe et al., 2018a). The supervised methods exploited 500 pairs of equivalent words that are not used in the evaluation data[9], and weakly supervised methods exploited pseudo bilingual pairs of words (auto seeds): the words with the same spellings among different languages were deemed as equivalent words. We trained the cross-lingual baselines and our model in each language pair.

As baselines of multilingual word embedding models, we used Chen and Cardie (2018) with or without auto seeds. We also compared our model against the cross-lingual baselines. While Chen and Cardie (2018) and our model jointly train multilingual word embeddings, the cross-lingual models independently map the word embeddings of German, Spanish, and French into the English embedding space. Regarding Artetxe et al. (2018b) and Artetxe et al. (2018a), we omitted the re-weighting, whitening, and normalisation processes in the multilingual experiments[10]. This

[6] For {Tamil, Turkish}-Japanese, we aligned the {Tamil, Turkish}-English dictionaries with the Japanese-English dictionary.

[7] https://github.com/jujbob/multilingual-models

[8] The code is at `https://code.google.com/archive/p/word2vec`, and the default algorithm is Continuous Bag of Words (CBOW) with its window size 5

[9] These 500 words were also extracted in the same way as explained in 4.2.

[10] To omit these processes, we used '–orthogonal' option

src	de		es		fr		ru		cs		fi		ja	
data size(tgt) Method	50k	1M	50k	1M	50k	1M	50k	1M	50k	1M	50k	1M	50k	1M
(weakly) supervised														
Artetxe et al. (2018b)+char	5.6	2.5	12.1	5.1	9.2	4.0	2.9	1.4	5.0	0.5	1.3	1.6	2.1	9.2
Artetxe et al. (2018a)+dict	9.6	**9.7**	15.0	19.7	13.3	19.5	**5.7**	**8.0**	5.5	**8.0**	**3.8**	**5.0**	6.1	11.2
Conneau et al. (2018)+dict	**11.1**	**9.7**	**18.0**	**20.4**	**19.2**	**20.7**	4.7	5.2	**7.1**	4.8	1.7	3.2	**7.5**	**18.7**
unsupervised														
Xu et al. (2018)	3.9	0.7	6.8	0.5	4.4	0.2	1.4	1.3	2.7	0.3	0.9	0.5	1.9	0.6
Artetxe et al. (2018b)	3.9	0.6	7.5	0.8	6.5	1.0	1.0	1.0	0.7	1.1	1.1	1.7	1.6	1.3
Conneau et al. (2018)	3.0	0.8	11.0	0.2	7.8	0.4	1.0	0.4	1.1	0.4	0.5	0.5	1.3	0.4
Conneau et al. (2018)+norm	2.1	0.7	11.3	0.7	9.2	0.3	0.7	0.3	0.6	0.2	0.6	0.3	1.7	0.5
OURS	**14.2**	**20.8**	**26.1**	**37.5**	**21.8**	**35.3**	**13.6**	**14.1**	**13.8**	**18.8**	**12.7**	**12.4**	**2.3**	**2.3**

Table 1: The precision p@5 of the cross-lingual word alignment task on the low-resource condition. We used 50k sentences for the source languages and either 50k or 1M sentences for the target language (English). The best scores among the (weakly) supervised or unsupervised methods are bold-faced, and the best scores of all the methods are underlined.

src-tgt Method	de-en	es-en	fr-en	ru-en	cs-en	fi-en	tr-ja	ta-ja	ja-en	se-fi	se-en
(weakly) supervised											
Artetxe et al. (2018b)+char	49.0	59.5	59.6	10.7	38.6	18.2	40.4	28.6	11.6	32.6	14.2
Artetxe et al. (2018a)+dict	35.6	49.7	49.4	38.5	38.5	28.3	25.8	**46.6**	24.5	42.9	20.2
Conneau et al. (2018)+dict	**53.6**	**66.8**	**67.6**	**53.1**	**54.0**	**43.4**	**41.1**	36.2	**34.1**	**43.8**	**32.5**
unsupervised											
Xu et al. (2018)	0.8	3.2	32.7	0.8	6.9	3.2	5.8	0.1	0.6	20.4	1.6
Artetxe et al. (2018b)	5.6	47.4	47.1	9.0	3.1	1.2	3.7	1.5	**1.8**	13.6	0.3
Conneau et al. (2018)	0.8	0.7	1.7	0.5	0.9	1.6	1.3	0.6	1.4	13.1	0.7
Conneau et al. (2018)+norm	0.7	2.5	0.6	0.6	0.3	0.2	0.1	1.3	0.9	23.2	0.2
OURS	**26.4**	**54.9**	**54.0**	**22.7**	**26.8**	**19.2**	**18.1**	**10.4**	**1.8**	**37.9**	**18.3**

Table 2: The precision p@5 of cross-lingual word alignment task on the different-domain condition. The best scores among the (weakly) supervised or unsupervised methods are bold-faced, and the best scores of all the methods are underlined.

is because these processes transform both source and target word embeddings, and that makes it impossible to map word embeddings of multiple languages into a single embedding space.

To implement these baselines, we used the code published by the authors[11,12,13](Conneau et al., 2018; Artetxe et al., 2018b; Xu et al., 2018)

4.4 Training Settings

In the cross-lingual and multilingual experiments, we trained our model among two and four languages, respectively. When the size of the source and target corpora were different, we conducted oversampling to generate the same number of mini-batches for source and target languages. We trained our model for 10 epochs with the mini-batch size 64, and stopped training when the training loss saturates (i.e., when the loss decreases by less than 1% compared to the previous epoch). For each iteration, our model alternately read mini-batches of each language and updated its parameters. We set the size of word embeddings as 300, and used two-layer LSTM networks for the forward and backward language models, respectively. We set the size of the hidden state as 300 and 1024 for the low-resource and different-domain conditions. Dropout (Srivastava et al., 2014) is ap-

in their code.

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

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

[13] https://github.com/xrc10/unsup-cross-lingual-embedding-transfer.

Method \ Condition	(a)	(b)	(c)
(weakly) supervised			
Artetxe et al. (2018b)+char	2.3	0.6	44.6
Artetxe et al. (2018a)+dict	5.6	**8.6**	37.2
Conneau et al. (2018)+dict	**6.4**	7.0	51.6
Chen and Cardie (2018)+char	5.2	2.8	**53.4**
unsupervised			
Xu et al. (2018)	1.1	0.0	0.0
Artetxe et al. (2018b)	0.9	0.0	5.2
Conneau et al. (2018)	1.3	0.0	0.0
Conneau et al. (2018)+norm	0.3	0.0	0.0
Chen and Cardie (2018)	1.0	0.0	3.0
OURS	**10.4**	**16.2**	**37.0**

Table 3: The precision p@5 of multilingual word alignment task on the three different conditions (a), (b), and (c) described in 4.1.2. The best scores among the (weakly) supervised or unsupervised methods are boldfaced, and the best scores of all the methods are underlined.

plied to the hidden state with a rate of 0.3. We used SGD (Bottou, 2010) as an optimiser with the learning rate 1.0. All of the parameters of our model including word embeddings were uniformly initialised in [-0.1, 0.1], and gradient clipping (Pascanu et al., 2013) was used with the clipping value 5.0. We included those words in vocabulary that were used at least 3, 5, and 20 times for 50k, 100k-250k, and 1M sentences in News Crawl and Wikipedia. For Europarl and SIKOR North Saami corpora, we set the threshold as 10. We fed the most 15,000 frequent words to train Xu et al. (2018) and the discriminator in Conneau et al. (2018).

As a preprocess, we tokenized the monolingual corpora using Moses toolkit[14] for European languages and Polyglot[15] for Tamil, Turkish and Japanese. We also lowercased all the corpora.

4.5 Results

4.5.1 Cross-lingual Word Embedding

Table 1 illustrates the results of the cross-lingual word alignment task under the low-resource condition. The methods with '+char' use character information to obtain a pseudo dictionary, and the ones with '+dict' use a gold dictionary that contains 500 pairs of words. The table shows that our model substantially outperforms the unsupervised baseline models in all of the language pairs. Our model also achieves better results than supervised methods except in the Japanese-English pair, which has different word order (SOV v.s SVO). Another interesting finding is that when the size of the target corpus increases from 50k to 1M sentences, our model improves its performance whereas the performance of the unsupervised baseline models drops substantially. For instance, when the size of the target corpus increases from 50k to 1M, Conneau et al. (2018) decreases the precision in Spanish-English from 11.0 to 0.2, while our model increases the precision from 26.1 to 37.5.

Table 2 shows the results on the different-domain condition. It shows that our method achieves better results overall than the unsupervised baseline models. The extremely poor performance of Conneau et al. (2018) under this condition is compatible with the results reported by Søgaard et al. (2018). Regarding the Japanese-English pair, none of the unsupervised methods including ours perform well, demonstrating that it is difficult to align languages without any supervision if the basic word order is different. Supervised methods, on the other hand, perform well in all the languages and outperform our model. This result indicates that even if domains of monolingual corpora are different across languages, the conventional approach of learning a linear transformation can be effective with (weak) bilingual supervision.

Impact of Data Size

To evaluate the effects of the data size on the model performances, we increased the size of both source and target corpora from 50k to 250k by 50k sentences. All of these sentences were extracted from News Crawl. Fig. 2 illustrates how p@5 changes depending on the data size. It shows that our model overall performs better than the baselines, especially among the distant language pairs such as Finnish-English. Although Artetxe et al. (2018b) report positive results on word alignment tasks between Finnish and English, our experiments show that their method requires much larger monolingual corpora such as Wikipedia on both the source and target sides to achieve good performance.

[14] https://github.com/moses-smt/mosesDecoder

[15] https://polyglot.readthedocs.io/en/latest/Tokenization.html

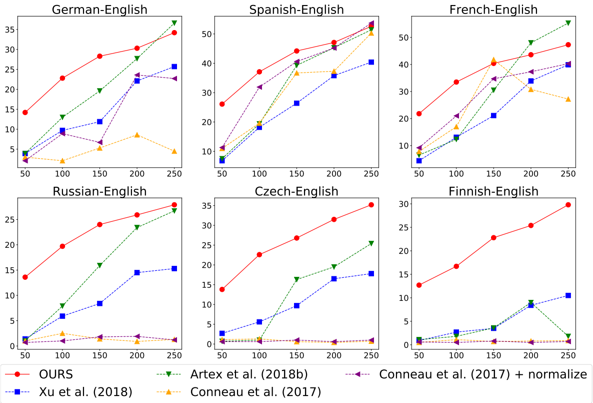

Figure 2: The change in p@5 achieved by the unsupervised methods on word alignment tasks. The x-axis denotes the number of sentences (thousand) in the source and target corpora, and the y-axis denotes the average precision p@5 over three runs for each method.

lang (src) / src-tgt	de	es	fr	ru	cs	fi	ja
same domain							
50k-50k	9.7	10.6	12.4	6.5	7.5	6.6	6.5
250k-250k	18.5	23.4	24.6	12.7	17.5	12.5	11.7
1M-1M	20.6	28.0	29.6	18.3	23.2	17.0	15.4
50k-1M	6.1	9.5	10.9	4.3	4.1	3.7	7.1
different domain							
1M-1M	15.7	19.5	22.2	17.9	19.5	15.2	13.6

Table 4: The ratio (%) of the monolingual word embeddings being roughly isomorphic across a source and target language (English). Each row describes the number of sentences in source and target corpora used to train word embeddings, and each column denotes the source language.

4.5.2 Multilingual Word Embedding

Table 3 describes the results under the three conditions described in 4.1.2. It shows that our model substantially outperforms the unsupervised and supervised baseline models under the low-resource conditions (a) and (b). As in the case of 4.5.1, when the size of the English corpus increases from 50k (a) to 1M (b), our model improves its performance while the unsupervised baselines perform worse. Under the different-domain condition (c), our model also achieves much better results than the unsupervised baselines, but cannot outperform supervised methods.

lang (src) / POS	de	es	fr	ru	cs	fi
ADJ	25.2	36.8	35.5	23.1	38.6	20.7
ADV	**68.8**	**82.6**	71.9	**82.6**	**81.6**	**66.2**
NOUN	24.5	53.8	51.1	13.2	16.4	9.2
VERB	16.1	66.7	**73.6**	34.4	34.9	19.7

Table 5: The ratio (%) of correctly matched POS tags using our model under the different-domain condition. For each language, the best and worst ratios among the four POS tags are bold-faced and underlined.

5 Analysis

5.1 Validation of Isomorphism

Our experiments show that our model substantially outperforms both supervised and unsupervised methods under the low-resource condition. We conjecture that this large improvement is owing to our unique approach of obtaining multilingual word embeddings; unlike the conventional approach, our method does not assume that word embedding spaces are approximately isomorphic across languages. In fact, when word embeddings are trained with small data, they should contain a lot of noises and are unlikely to be isomorphic across languages. This suggests that it would be extremely difficult to learn a linear mapping across languages using the existing unsupervised methods.

To verify this hypothesis, we investigated how likely monolingual word embeddings were more or less isomorphic across languages. For each pair of a language ℓ and English, we sampled 10 pairs of equivalent words from a bilingual dictionary and built non-directed adjacency matrices of the nearest neighbour graphs $G(\ell)$ and $G(en)$ independently. Then, we conducted an element-wise comparison of the two matrices and deemed them as roughly isomorphic if more than 80% of the elements are the same. Table 4 shows how often the graphs were roughly isomorphic over 1,000 samples. The row indicates the size of the source and target corpora. It clearly shows that monolingual corpora trained on small data (i.e. 50k sentences) are far from being isomorphic between any language pair, and the linguistically distant languages such as Finnish-English and Japanese-English are less isomorphic than close languages. This result clearly explains why the existing unsupervised methods do not perform well on the low-resource condition, or among distant language pairs. Another intriguing finding is that word embeddings trained with 50k and 1M sentences in a source and target languages are overall less isomorphic than those trained with 50k source and target sentences. This result explains why the performance of the unsupervised baseline methods decreases given the additional target data in 4.5.1 and 4.5.2. Our method, on the other hand, can effectively utilise the additional data to improve its performance, demonstrating its robustness under the low-resource condition.

5.2 POS tags of Matched Words

To analyse the performance of our model, we checked Part-of-Speech (POS) tags of the English words used in the word alignment task and investigated what kind of words were correctly matched by our model. Since a word is given without any context in the word alignment task and it is not possible to infer its POS tag, we assigned to each word its most frequent POS tag in Brown Corpus (Kucera and Francis, 1967). For instance, since 'damage' is used as a noun more often than as a verb in Brown Corpus, we define its POS tag as 'noun'. Table 5 shows p@5 of the word alignment task grouped by the four major POS tags, namely adjective, adverb, verb, and noun[16]. It clearly indicates that an adverb can be easily matched in every language pair. This would be because there are less adverbs than other tags in the evaluation data, and also because there are common word order rules about an adverb among all the languages: an adverb usually comes before an adjective to modify it, and when modifying a verb, it comes either before or after it. Refer to the Appendix B for the statistics regarding word order in each language. Among French, Spanish and English, the matching accuracy of a noun and verb is very high, and their word order is in fact very similar; as shown in the Appendix B, the basic word order of these languages is strictly subject-verb-object, and that makes it easy to align words among them. On the other hand, the word order between a noun and adjective is very different among these languages, explaining why the precision of matching adjectives is lower than the other tags. As for the other languages, they have more flexible word order than English and that makes it difficult to align words across languages. For instance, in German, Russian and Czech a subject sometimes comes after a verb, and in German and Finnish an object can come before a verb. These findings clearly indicate that our model employs sequential similarities among different languages to obtain multilingual word embeddings without any supervision.

6 Conclusion

In this paper, we proposed a new unsupervised multilingual word embedding approach. Whereas conventional methods aim to map pre-trained word embeddings into a common space, ours jointly generates multilingual word embeddings by extracting a common language structure among multiple languages. Our experiments on word alignment tasks have demonstrated that our proposed model substantially outperforms the existing cross-lingual and multilingual unsupervised models under resource-poor conditions, namely when only small data are available or when domains of corpora are different across languages. Under the first condition, our model even outperforms supervised methods trained with 500 bilingual pairs of words. By analysing the nearest neighbour graphs of monolingual word embeddings, we have verified that word embeddings are far from being isomorphic when they are trained on small data, explaining why existing unsupervised methods did not perform well on the low-resource condition. We have also found that the performance of our model is closely related to word order rules, and our model can align words very well when they are used in a similar order across different languages. Our future work is to exploit character and subword information in our model and see how those information affect the performance in each language pair. It would be also interesting to investigate how our approach compares to the baselines given a large amount of data such as Wikipedia.

7 Acknowledgement

We are grateful to all the anonymous reviewers for their insightful comments and advice.

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the*

[16] When there are X nouns and Y of them are matched correctly in the alignment task, the ratio is $\frac{100Y}{X}\%$

European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 937–947. Association for Computational Linguistics.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised hyperalignment for multilingual word embeddings. *CoRR*, abs/1811.01124.

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

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

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France. Springer.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium. Association for Computational Linguistics.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

H. Kucera and W. N. Francis. 1967. *Computational analysis of present-day American English*. Brown University Press.

KyungTae Lim, Niko Partanen, and Thierry Poibeau. 2018. Multilingual Dependency Parsing for Low-Resource Languages: Case Studies on North Saami and Komi-Zyrian. Miyazaki, Japan. ELRA.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (Workshop)*.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*.

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

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1310–1318.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

V Rudramurthy, Mitesh M. Khapra, and Pushpak Bhattacharyya. 2016. Sharing network parameters for crosslingual named entity recognition. *CoRR*, abs/1607.00198.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*.

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. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 1006–1011. Association for Computational Linguistics.

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

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

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

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398. Association for Computational Linguistics.

dep-head(rel)	en	de	es	fr	ru	cs	fi
N-V(nsubj)	93.7	77.5	89.0	95.4	80.4	77.1	88.2
N-V(obj)	0.7	56.3	0.5	0.2	4.0	10.0	20.8
ADJ-N	98.9	99.9	30.3	30.9	99.3	93.9	100.0
ADV-ADJ	98.3	93.6	95.0	99.2	96.6	94.9	98.8
ADV-V	75.6	65.8	68.7	61.2	79.1	80.4	47.9

Table 6: The ratio (%) of a dependent being put before its head. N, V, ADJ, and ADV denote noun, verb, adjective and adverb, respectively. The dependency relation of ADJ-N is amod, and the one of ADV-ADJ and ADV-V is advmod. Refer to the download page of PUD for the definition of the dependency relations.

A Visualisation

Figure 3 visualises the multilingual word embeddings obtained by our model and (Chen and Cardie, 2018) under the low-resource condition. It shows the most frequent 1000 words in Spanish, French, German and English. The figure clearly shows that the word embeddings obtained by (Chen and Cardie, 2018) form some clusters based on their languages. In particular, many of the German words are mapped near the centre of the figure and make a large cluster. On the other hand, the word embeddings trained by our model are not clustered by language, indicating that our model successfully maps word embeddings into a common space.

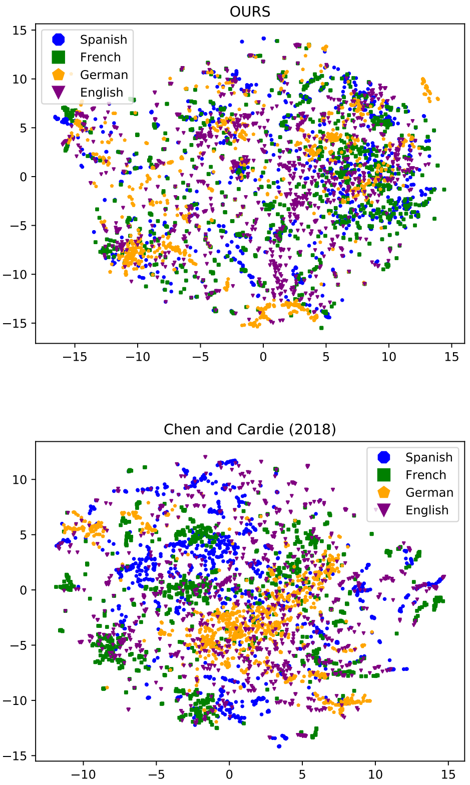

Figure 3: Scatter plot of multilingual word embeddings of French, English, German and Spanish obtained by our model and Chen and Cardie (2018) under the low-resource condition. The embeddings are reduced to 2D using tSNE (van der Maaten and Hinton, 2008).

B Word Order

To obtain statistics about word order rules in each language, we used Parallel Universal Dependencies (PUD) treebanks[17]. PUD contains 1000 parallel sentences aligned among 18 languages, and those sentences are annotated morphologically and syntactically according to Google universal annotation guidelines. Since these sentences are aligned among all the languages, it is possible to compare the syntactical differences across languages.

Table 6 shows the ratio of a dependent being put before its head in PUD treebanks in each language. As can be seen, the word order of ADV-

[17] available at http://universaldependencies.org/

ADJ (advmod) is very similar among all the language pairs: an adverb is put before an adverb to modify it. The order of ADV-V (advmod) is rather flexible regardless of language, indicating that an adverb can modify a verb from either left or right. These common word order rules of adverbs explain why our model successfully matched adverbs very well in every language pair. The table also indicates that the word order of N-V is very similar among English, Spanish and French and the basic word order is strictly subject-verb-object. This explains why our model performed well overall among these languages. However, the word order of ADJ-N is significantly different among these languages, and that would lead to the low performance of our model in matching adjectives.