EasyChair Preprint

№ 2254

Unsupervised Cross-lingual Word Embeddings Based on Subword Alignment

Jin Sakuma and Naoki Yoshinaga

December 25, 2019

Unsupervised Cross-lingual Word Embeddings Based on Subword Alignment

Jin Sakuma[1] and Naoki Yoshinaga[2]

[1] The University of Tokyo
jsakuma@tkl.iis.u-tokyo.ac.jp
[2] Institute of Industrial Science, the University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Abstract. Cross-lingual word embeddings are crucial building blocks for multilingual models, and recent studies indicate that they are obtainable without any bilingual resources [4, 8]. However, the reported experimental results (replicated in this study) indicate that the performance of such cross-lingual word embeddings degrades on distant language pairs such as English-Japanese and English-Finnish. In this paper, we propose an unsupervised method to obtain cross-lingual word embeddings by exploiting unambiguously-translatable word pairs such as loanwords and named entities as a dictionary to induce mappings. Given an initial bilingual dictionary (obtained in an unsupervised manner), our method learns a subword alignment model to extract unambiguously translatable word pairs whose surfaces are alignable with each other. We then employ the bilingual dictionary refined with the subword alignment model to induce accurate cross-lingual word embeddings. Experimental results indicate that cross-lingual word embeddings obtained with our method were more accurate than those obtained by the state-of-the-art method, especially on distant language pairs.

1 Introduction

Various tasks in natural language processing (NLP) have undergone significant improvements in accuracy by using neural networks that are trained on massive corpora annotated for the given task in the target language. However, preparing such corpora for every combination of tasks and languages is unrealistic, and as a result, the models in many resource-poor languages have the degraded accuracy. To mitigate this problem, several researchers investigated methods to enable cross-lingual transfer of a model (hereafter, multilingual model) which are trained in a resource-rich language (hereafter, source language) and can be applied to another resource-poor language (hereafter, target language) [6, 7, 16]. Most of these researches exploit cross-lingual word embeddings, which are vector representations of words in the same semantic space across languages, to absorb the difference in the vocabularies among languages.

Although existing studies [4, 8] have successfully obtained cross-lingual word embeddings, our experimental results confirmed that these methods work poorly for distant languages pairs (§ 4.2). This is partly because that word translation tends to be ambiguous for distant language pairs. The polysemous words in distant languages are likely to share only a part of their senses, and the remaining senses are irrelevant to each other.

For example, an English word "moon" has multiple translations in Japanese such as "月 (The moon)," and "衛星 (satellite)," while "月" has multiple translations in English such as *Monday* and *month*, which are not included in the meaning of moon. The inclusion of such polysemous words into the bilingual dictionary prevents us from inducing reliable cross-lingual word embeddings.

To mitigate this problem of ambiguous word-to-word correspondences in distant language pairs, we take advantage of words that have surface correspondences such as loanwords and named entities to induce cross-lingual word embeddings. We assume that such word pairs with surface correspondences are likely to be unambiguously translatable with each other since those words originally came from the other language. To find such words from the bilingual dictionary, we exploit subword alignment to extract *well-aligned* word pairs.

Given an initial bilingual dictionary, we train a subword alignment model [13] to assign an alignment score to each word pair in the dictionary. We then extract word pairs with greater alignment scores in the dictionary to create an unambiguously translatable bilingual dictionary as they are expected to include loanwords or named entities. When combined with the unsupervised method of bilingual dictionary induction [4], our method can work in a fully unsupervised manner and does not rely on any cross-lingual resources such as a bilingual dictionary or a parallel corpus.

The contributions of this paper are as follows:

- We proposed a novel method to obtain cross-lingual word embeddings that exploits subword alignment.
- Our method advanced the state-of-the-art for the task of inducing cross-lingual word embeddings for distant language pairs without supervision.
- We experimentally confirmed that the quality of cross-lingual word embeddings obtained through an existing method [4] is degraded in distant language pairs.
- We evaluated the performance of cross-lingual word embeddings when using non-comparable corpora obtained from Twitter to learn monolingual word embeddings.

The structure of this paper is as follows. In § 2, we introduce existing methods for inducing cross-lingual word embeddings and related studies that utilized subword information for multilingual models. In § 3, we propose a novel method to obtain cross-lingual word embeddings by exploiting subword alignment. In § 4, we conduct a series of experiments to evaluate our method and understand its characteristics. Finally, we will summarize our work in § 5.

2 Related Work

In what follows, we first introduce existing methods to obtain cross-lingual word embeddings (§ 2.1). We then discuss other studies that exploit subword information for multilingual NLP (§ 2.2).

2.1 Cross-lingual word embeddings

Most of the existing methods for learning cross-lingual word embeddings first obtain monolingual word embeddings for each language and then learn a mapping between

the obtained embeddings. Early studies for these methods exploit hand-built bilingual resources such as a bilingual dictionary and parallel corpus to induce the mappings [1, 3, 9, 10, 15, 18].

It is impractical to prepare bilingual resources for every language pair since it requires a costly manual annotation, and thus some researches focus on obtaining cross-lingual word embeddings with minimal supervision. Artetxe et al. [2] successfully obtained cross-lingual word embeddings from 25 words pairs or numerals by self-learning framework which alternates between inducing the dictionary and training the mapping.

Recently, there have been several successful attempts to obtain cross-lingual word embedding in a fully unsupervised manner. Smith et al. [17] developed an unsupervised method that utilizes word pairs with the same exact character string as a bilingual dictionary. Conneau et al. [8] exploited adversarial learning to obtain cross-lingual word embeddings without any cross-lingual supervision. Artetxe et al. [4] enhanced their self-learning framework with an unsupervised initialization strategy and a robust learning method. These unsupervised methods exhibited comparable or even better performance in similar language pairs against the ones based on bilingual resources; in other words, hand-built cross-lingual resources are not always optimal for obtaining cross-lingual word embeddings.

Although most of the above studies study mainly work on similar language pairs in European languages, we should rather work on unsupervised cross-lingual word embeddings between resource-rich languages (such as English and European languages) and resource-poor languages that are distant from English, considering the target of multilingual models. However, the performance of the above unsupervised models is still limited for distant language pairs such as English and Japanese as we will later demonstrate in § 4.

In this work, we focus on such distant language pairs and improve cross-lingual word embeddings by utilizing subword alignment under the same unsupervised settings as [4]. We compare our method with the state of the art method [4] that achieved the best performance for distant languages pairs such as English and Finish, to advance the state-of-the-art.

2.2 Exploiting subwords for cross-lingual transfer

There are a few studies that exploit subword information to enable cross-lingual transfer of a model. Several studies use character-level embeddings shared across languages for cross-lingual POS tagging [12, 19]. Ishiwatari et al. [11] learn a cross-lingual projection of word representation by exploiting subword information in addition to a hand-built bilingual dictionary. Another study induces cross-lingual word embeddings by sharing subword information by directly applying Subword-Information Skip-gram (SISG) [5] on a joint corpus of two languages which are then used as a building block for unsupervised machine translation [14].

These methods are, however, not applicable to various distant language pairs especially when they have different character sets. We overcome this issue by learning a subword alignment model which is capable of inducing relationships between different character sets. Our method will contribute to making the above subword-based methods applicable to distant language pairs with different character sets.

3 Proposal

Here, we explain the details of our method to obtain cross-lingual word embeddings of two languages by exploiting subword alignment. Given an initial bilingual dictionary, our method improves the quality of the dictionary for inducing bilingual word embeddings by filtering ambiguous word translation such as moon (the moon, satellite, moonlight, etc.) and 月 (Japanese words for the moon, Monday, month, etc.). We first train a subword alignment model to compute an alignment score for each word pair in the dictionary. We then collect word pairs with high alignment scores to construct a refined bilingual dictionary which we expect to contain mostly unambiguously translatable word pairs. The refined bilingual dictionary is finally used to train cross-lingual word embeddings using the existing state-of-the-art supervised method [3].

We hereafter explain each step of our method in detail:

Step 0: Preparing initial dictionary First, we need to prepare an initial bilingual dictionary. To address a common situation where no hand-built bilingual resources are available, we expect to induce a bilingual dictionary in an unsupervised manner [4]. Of course a hand-built bilingual dictionary can be adopted, if any.

Step 1: Learning subword alignment model Given the initial bilingual dictionary. we train a subword alignment model that computes the likelihood of character-level alignment of word pairs in the dictionary. For this purpose, we exploit a many-to-many alignment method [13] that is capable of aligning two sequences of symbols (words) for any language pair. We expect this model to learn how words are imported from one language to another.

Suppose that $D_{init} = \{(x_1, y_1), \cdots, (x_N, y_N)\}$ is the initial bilingual dictionary, where x_i and y_i are words (sequence of characters) in the source and the target languages. For each word pair (x_i, y_i), we want to find alignment $\mathbf{u}$ that is most likely to happen.

$$\hat{\mathbf{u}} = \underset{\mathbf{u} \in \mathcal{U}_{(\mathbf{x_i}, \mathbf{y_i})}}{\arg\max}\, P(\mathbf{u}|(\mathbf{x_i}, \mathbf{y_i}))$$

where $\mathcal{U}_{(x_i, y_i)}$ is the set of all possible alignment of x_i and y_i. This model is trained by an Expectation-Maximization algorithm.

Step 2: Filtering the initial bilingual dictionary Now, we filter the bilingual dictionary induced in Step 1 so that we can obtain word pairs that have less ambiguity in mutual translation. For each word pair (x_i, y_i), we compute the best character-level alignment $\hat{\mathbf{u}}$ and its alignment score, $\log P(\hat{\mathbf{u}}|(x_i, y_i))$. We extract word pairs with alignment scores higher than a threshold to construct the refined bilingual dictionary $D_{refined}$.

An issue here is that we may not have any development set to tune the threshold because we want to maximize our system's applicability by not relying on any hand-built bilingual resources. To find the best threshold for the alignment score, we take 100 word pairs in the refined dictionary with the highest alignment scores to be a development set, and we donate the remaining bilingual dictionary as $D'_{refiend}$. We created multiple refined dictionaries with different thresholds, and the resulting cross-lingual word embeddings are evaluated on bilingual dictionary induction task using the development set. The cross-lingual word embeddings with the best performance on the development set are used in the evaluation.

Step 3: Training cross-lingual word embeddings We now train cross-lingual word embeddings from the reliably subword-aligned (which we expect to be unambiguously translatable) bilingual dictionary obtained in Step 2. We employ an existing method for supervised training of cross-lingual word embeddings [3].

Given word embeddings of the source and the target languages, X and Y, and the refined bilingual dictionary $D'_{refined}$, this method trains two mappings W_x and W_y so that the mapped embeddings XW_x and YW_y are in the same semantic space by minimizing the following objective.

$$\left(\hat{W}_x, \hat{W}_y\right) = \underset{W_x, W_y}{\arg\max} \sum_{i,j \in D_{refined}} (X_i W_x) \cdot (Y_j W_y)$$

This objective ensures word pairs in the bilingual dictionary $D'_{refined}$ become similar after mapping. To enhance the quality of cross-lingual word embeddings, embeddings are normalized and whitened so that different components have unit variance and be uncorrelated before learning mappings and de-whitened to restore the original variance after. Like many other methods [2,8,17], the mappings are constrained to be orthogonal. The reader will refer to the original paper for the details.

4 Evaluation

To examine the effect of exploiting subword alignments and gain a profound understanding of our method, we conduct experiments on obtaining cross-lingual word embeddings for various language pairs. Following existing studies [2, 3, 8], we used the bilingual lexicon induction task for evaluation. We first conduct a detailed evaluation in four language pairs including two distant language pairs, English-Japanese (en-ja) and English-Finnish (en-fi) and two similar language pairs, English-Spanish (en-es) and English-Italian (en-it) (§ 4.2). We then evaluate our method in eight additional language pairs to evaluate their performances in various situations (§ 4.3). To further evaluate the applicability of our method in various situations, we conduct experiments on monolingual word embeddings trained on the Twitter corpus (§ 4.4). Finally, we conduct a qualitative analysis of the refined bilingual dictionaries obtained in Step 2 (§ 3).

4.1 Settings

In the following, we explain the details of the experimental settings. We first introduce the bilingual lexicon induction for evaluation task, next detail methods for comparison, and then explain how to obtain monolingual word embeddings.

Bilingual lexicon induction Bilingual lexicon induction is a task to predict the translation of a word in the source language in the target language. Given a word in the source language, we take the closest word in the target language, and if the word is in the set of translations of the source word in the ground truth bilingual dictionary, we consider it to be correct. For all evaluations, we used the test portion of MUSE bilingual dictionary[3] as the ground truth dictionary which are used in previous studies [4, 8].

[3] `https://github.com/facebookresearch/MUSE`

Methods for comparison In order to evaluate the impact of exploiting subword alignment to filter out a bilingual dictionary used to induce cross-lingual word embeddings [3], we compare six methods that differ in how to prepare the bilingual dictionary for inducing cross-lingual word embeddings [3].

Method #1 (Unsupervised) This unsupervised baseline method is [4] that iteratively repeats bilingual dictionary induction and learning cross-lingual word embeddings.
Method #2 (Unsup. w/ CSLS filtering) This method filters the bilingual dictionary used in the final iteration of Method #1 using CSLS similarities of the word pairs.
Method #3 (Unsup. w/ our filtering) Our method (§ 3) filters the bilingual dictionary used in the final iteration of Method #1 using subword alignment scores.
Method #4 (Supervised) This supervised baseline method [3] uses the training portion of MUSE bilingual dictionary.
Method #5 (Sup. w/ our filtering) Our method (§ 3) filters the bilingual dictionary used in Method #4 using subword alignment scores.
Method #6 (Sup. + Unsup. w/ our filtering) This method combines the hand-built bilingual dictionary used in Method #4 and the bilingual dictionary obtained by Method #3 (with a different threshold to alignment scores).

For the unsupervised methods with filtering (**Method #2** and **#3**), we kept 100 word pairs of the induced initial dictionary with the highest CSLS similarities in the development set to tune the filtering threshold of CSLS similarity and alignment scores, respectively, and the remaining word pairs are used as the training set. For the supervised methods with filtering (**Method #5** and **#6**), we randomly sampled 500 word pairs from the bilingual dictionary as the development set to tune the filtering threshold of the alignment scores, and the remaining word pairs are used as the training set.

Monolingual word embeddings All of the above methods require monolingual word embeddings of the source and target languages to obtain cross-lingual word embeddings. For this purpose, we used pre-trained word embeddings available online[4] for all languages, which are obtained by applying subword-information skip-gram (SISG) [5] to the Wikipedia dump files,[5] except for Japanese[6] and experiments on Twitter corpora (detailed in § 4.4). For Japanese and experiments on the Twitter corpora, we used the official implementation of SISG[7] to obtain word embeddings from a Japanese Wikipedia dump file of 2018-11 (tokenized by MeCab v0.996[8]) and the Twitter corpora. In all of our experiments, we take the 200,000 most frequent words as our vocabulary for each language.

Implementation For character-level many-to-many alignment in Step 1 of our method, we used mpaligner[9] version 0.97. To learn a mapping across languages in Step 0 of **Method #1**, **#2**, **#3** and **#6** and Step 2, we used the official implementation[10] of the original papers [3, 4] with the default hyperparameters.

[4] https://fasttext.cc/docs/en/pretrained-vectors.html
[5] https://dumps.wikimedia.org/
[6] Japanese pre-trained embeddings available online were broken.
[7] https://github.com/facebookresearch/fastText
[8] http://taku910.github.io/mecab/
[9] https://osdn.net/projects/mpaligner/
[10] https://github.com/artetxem/vecmap

Table 1: Results of bilingual lexicon induction. Accuracy marked with $*$ was significantly better than the unsupervised (#1) and supervised (#4) baselines ($p < 0.05$ assessed by Wilcoxon signed-rank test).

	distant		similar	
Method	**en-ja**	**en-fi**	**en-es**	**en-it**
#1 Unsup.	0.4573	0.4393	0.8086	0.7713
#2 Unsup. w/ CSLS filtering	0.4440	0.4400	0.8000	0.7673
#3 Unsup. w/ our filtering	**0.4874***	**0.4547***	**0.8087**	**0.7787**
#4 Sup.	0.5175	0.4373	0.7940	0.7587
#5 Sup. w/ our filtering	0.4944	0.4320	0.7913	0.7580
#6 Sup. + Unsup. w/ our filtering	**0.5210**	**0.4766***	**0.8033**	**0.7686**

4.2 Detailed evaluation in four language pairs

Table 1 shows the performance of our methods (**Methods #3**, **#5**, and **#6**) and the baseline methods (**Method #1**, **#2**, and **#4**) in four language pairs: English-Japanese (en-ja), English-Finnish (en-fi), English-Spanish (en-es), and English-Italian (en-it).

Comparison with the unsupervised baseline (Method #1 vs. #3) Our unsupervised method (**Method #3**) outperforms the unsupervised baseline method (**Method #1**) in all of the four languages. Furthermore, the differences in the accuracies were statistically significant for distant language pairs: en-ja and en-fi. From these results, we confirmed the effectiveness of our method, especially for distant language pairs.

Comparison with the alternative filtering method (Method #2 vs. #3) Next, we examine if we genuinely need subword alignment, or if other simple methods of filtering also yield similar results. **Method #2** filtered unreliable word translation by CSLS similarity scores used in Step 1 (**Method #1**) instead of the alignment scores. This filtering does not consider the ambiguity of word translation, although it will yield a smaller but higher quality bilingual dictionary since it keeps only word pairs with high confidence.

Our method (**Method #3**) outperforms the alternative filtering method for all language pairs. Furthermore, the CSLS filtering (**Method #2**) degraded the accuracy of the unsupervised baseline (**Method #1**). We thus conclude that subword alignment provides useful information to improve the quality of cross-lingual word embeddings.

Evaluation of supervised methods (Method #4 #5, and #6) Occasionally, a hand-built bilingual dictionary is available to obtain cross-lingual word embeddings. Here, we consider what method is suitable in such a situation. We compare three supervised methods including the baseline [3] (**Method #4**), and two modified versions of our method (**Method #5** and **#6**).

Method #6, the combination of the hand-built dictionary and the automatically-induced dictionary further filtered by subword alignment, yielded the best performance among the supervised methods (**Method #4**, **#5**, **#6**) for all of the language pairs. **Method #6** is even better than the best-performing unsupervised method (**Method #3**) for the two distant language pairs. For the two similar language pairs, we found the best unsupervised method (**Method #3**) outperforms the supervised methods pairs.

Table 2: Results of bilingual lexicon induction in additional eight language pairs. Accuracy marked with * was statistically better than the other($p < 0.05$ assessed by Wilcoxon signed-rank test).

Method	en-da	en-de	en-fr	en-nl	en-pt	en-sv	en-tr	en-fa
#1 Unsup.	0.5567	0.7327	**0.8040**	0.7333	0.7853	0.6040	0.4827	**0.3147**
#3 Unsup. w/ our filtering	**0.6100***	**0.7373**	0.8013	**0.7347**	**0.8020***	**0.6233***	**0.4833**	0.3127

Table 3: Statistics of Twitter corpora.

Lang.	# tweets (m)	Ave. # tokens
English (en)	193	14.18
Japanese (ja)	117	19.32
Finnish (fi)	26	17.01
Spanish (es)	43	14.62
Italian (it)	93	16.47

4.3 Evaluation in various language pairs

To evaluate our method in various situations, we compare our **Method #3** with unsupervised baseline method [4] (**Method #1**) in eight additional language pairs: English-Danish (en-da), English-German (en-de), English-French (en-fr), English-Dutch (en-nl), English-Portuguese (en-pt), English-Swedish (en-sv), English-Turkish (en-tr), and English-Persian (en-fa).

The result is shown in Table 2. Our method (**Method #3**) significantly outperformed the unsupervised baseline (**Method #1**) in three of eight language pairs (en-da, en-pt, and en-sv), while it is comparable in other language pairs. This result confirms the applicability of our method in various language pairs.

4.4 Evaluation on Twitter corpus

The Wikipedia corpora we used to induce monolingual word embeddings in the experiments are comparable corpora rather than independent monolingual corpora since many articles are on the same topics across languages. As pointed out by the existing study [4], such corpora may expose strong cross-lingual signal which is not obtainable in a strictly unsupervised situation.

To evaluate our method in a more realistic situation, we conducted experiments on word embeddings obtained from Twitter corpora. We obtained tweets (excluding retweets) in August of 2017 in English, Japanese, Finnish, Spanish, and Italian. User IDs starting from "@" are replaced with a special token, and all URLs are removed. We then tokenized the tweets using MeCab v0.996[11] for Japanese, and NLTK[12] for Finnish, Spanish, and Italian. Table 3 summarizes the statistics of the resulting corpora.

[11] `http://taku910.github.io/mecab/`
[12] `https://www.nltk.org/api/nltk.tokenize.html`

Table 4: Results on bilingual lexicon induction using Twitter corpora. Accuracy marked with * was significantly better than the other ($p < 0.05$ assessed by Wilcoxon signed-rank test).

	distant		similar	
Method	**en-ja**	**en-fi**	**en-es**	**en-it**
#1 Unsup.	**0.2898***	0.7831	0.5223	0.4386
#3 Unsup. w/ our filtering	0.2810	**0.7908***	**0.5534***	**0.4428***

The results of bilingual dictionary induction are shown in Table 4. Note that the accuracy in this table are not comparable to those in Table 1 since they are evaluated only when monolingual word embeddings are available for words and their translations. Among three of four language pairs tested, our method (**Method #3**) outperformed the unsupervised baseline method (**Method #1**).

Interestingly, we obtained significant improvements on similar language pairs (en-es and en-it) in this experiments with Twitter corpora, while we could not obtain significant improvement on similar language pairs when we use pre-trained monolingual word embeddings obtained from Wikipedia (Table 1). This is probably because the use of monolingual word embeddings obtained from non-comparable corpora increased the problem of ambiguous word translation even in similar language pairs when [4] is used to induced the initial bilingual dictionary. Even if two words in similar language pairs share most of the meanings, it does not guarantee their embeddings have good correspondences since those words can refer to different meanings in non-comparable corpora. This also deteriorates the quality of the development set to tune the threshold for our method, which degraded the accuracy on en-ja (the most distant language pairs).

4.5 Qualitative analysis

Finally, we confirm if our filtering method correctly obtains unambiguously-translatable word pairs such as loanwords and named entities. From the refined bilingual dictionary obtained in Step 2 (§ 3), we present top-10 word pairs with the highest alignment scores excluding ones with the same character string in Table 5. We also show the alignment score ranking including word pairs with the same character strings.

We can see that we successfully obtained loanword pairs such as cost-コスト in English-Japanese, camera-kamera in English-Finnish, and international-internacional in English-Spanish, and named entities such as india-intia in English-Finnish, and americans-americani in English-Italian. Also, we found that the model correctly associates suffix "-s" in English with suffix "-i" in Italian which both indicates plural and "c" in English with "k" in Finnish.

Through these observations, we found that even though the bilingual dictionary induced in an unsupervised manner contains incorrectly translated word pairs and word pairs without perfectly-aligned subwords, the obtained many-to-many alignment model correctly models transliteration and correspondences between subwords that have the same grammatical functionalities.

Table 5: Word pairs in the bilingual dictionaries refined by using subword aignment.

rank	English Japanese
1	chart チャート *(tya a to)*
2	demonstration デモンストレーション *(de mo n su to re e sho n)*
3	plantation プランテーション *(pu ra n te e sho n)*
4	sparta スパルタ *(su pa ru ta)*
5	elf エルフ *(e ru hu)*
6	scrap スクラップ *(su ku ra ppu)*
7	ana アナ *(a na)*
8	timing タイミング *(ta i mi n gu)*
9	scandal スキャンダル *(su kya n da ru)*
10	brest ブレスト *(bu re su to)*

(a) English-Japanese

rank	English Finnish
68	croatia kroatia
138	constantin konstantin
139	israelis israelin
196	india intia
213	socrates sokrates
227	camera kamera
286	macedonian makedonian
326	atlantic atlantin
332	tina nina
336	caucasian kaukasian

(b) English-Finnish

rank	English Spanish
323	international internacional
487	secretaries secretarios
496	territories territorios
591	mercenaries mercenarios
606	initial inicial
628	rational racional
653	residential residencial
666	national nacional
702	narrator narrador
705	salaries salarios

(c) English-Spanish

rank	English Italian
439	italians italiani
453	terrorists terroristi
502	errors errori
532	senators senatori
558	arrests arresti
616	tensions tensioni
625	americans americani
657	assassins assassini
658	continents continenti
688	aliens alieni

(d) English-Italian

5 Conclusions

In this paper, we proposed an unsupervised method to refine a bilingual dictionary for inducing accurate cross-lingual word embeddings. Our method exploits subword alignment to extract unambiguously translatable word pairs from a given bilingual dictionary. Experimental results confirmed that our method advanced the state-of-the-art for the task of inducing cross-lingual word embeddings, especially for distant language pairs and when non-comparable corpora are used for obtaining monolingual word embeddings. Our method successfully identified loanwords and named entities that are expected to be helpful to obtain cross-lingual word embeddings.

Although our method improved the quality of unsupervised cross-lingual word embeddings of distant language pairs, the performance is still not comparable to that of similar language pairs. We believe that this is due to the difference in grammars (word order) and word segmentation across languages, and it remains to be solved in the future, to further improve cross-lingual word embeddings for distant language pairs.

Acknowledgements

This work was partially supported by Commissioned Research (201) of the National Institute of Information and Communications Technology of Japan.

References

1. Artetxe, M., Labaka, G., Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: Conference on Empirical Methods in Natural Language Processing (2016)
2. Artetxe, M., Labaka, G., Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data. In: Annual Meeting of the Association for Computational Linguistics (2017)
3. Artetxe, M., Labaka, G., Agirre, E.: Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In: Association for the Advancement of Artificial Intelligence (2018)
4. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Annual Meeting of the Association for Computational Linguistics (2018)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics (2017)
6. Can, E.F., Ezen-Can, A., Can, F.: Multilingual sentiment analysis: An rnn-based framework for limited data. CoRR (2018)
7. Chen, X., Sun, Y., Ben, A.: Adversarial deep avaraging networks for cross-lingual sentiment classification. In: Conference on Empirical Methods in Natural Language Processing (2017)
8. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: International Conference on Learning Representation (2018)
9. Dinu, G., Baroni, M.: Improving zero-shot learning by mitigating the hubness problem. In: International Conference on Learning Representation (2015)
10. Faruqui, M., Dyer, C.: Improving vector space word representations using multilingual correlation. In: Conference of the European Chapter of the Association for Computational Linguistics (2014)
11. Ishiwatari, S., Kaji, N., Yoshinaga, N., Toyoda, M., Kitsurekawa, M.: Accurate cross-lingual projection between count-based word vectors by exploiting translatable context pairs. In: Conference on Computational Language Learning (2015)
12. Kim, J.K., Kim, Y.B., Sarikaya, R., Fosler-Lussier, E.: Cross-lingual transfer learning for pos tagging without cross-lingual resources. In: Conference on Empirical Methods in Natural Language Processing (2017)
13. Kubo, K., Kawanami, H., Saruwatari, H., Shikano, K.: Unconstrained manyto-many alignment for automatic pronunciation annotation. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (2011)
14. Lample, G., Ott, M., Conneau, A., Denoyer, L., Ranzato, M.: Phrase-based & neural unsupervised machine translation. In: Conference on Empirical Methods in Natural Language Processing (2018)
15. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv:1309.4168 (2013)
16. Pappas, N., Popescu-Belis, A.: Multilingual hierarchical attention networks for document classification. In: International Joint Conference on Natural Language Processing (2017)
17. Smith, S.L., Turban, D.H.P., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: International Conference on Learning Representations (2017)

18. Xing, C., Wang, D., Liu, C., Lin, Y.: Normalized word embedding and orthogonal transform for bilingual word translation. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2015)
19. Yang, Z., Salakhutdinov, R., Cohen, W.W.: Transfer learning for sequence tagging with hierarchical recurrent networks (2017)