Improving Japanese-English Bilingual Mapping of Word Embeddings based on Language Specificity

Yuting Song*, Biligsaikhan Batjargal† and Akira Maeda‡
* *Research Organization of Science and Technology, Ritsumeikan University, Kusatsu, Japan 525–8577*
Email: ytsong@gst.ritsumei.ac.jp
† *Kinugasa Research Organization, Ritsumeikan University, Kyoto, Japan 603–8577*
Email: biligee@fc.ritsumei.ac.jp
‡ *College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan 525–8577*
Email: amaeda@is.ritsumei.ac.jp

***Abstract*—Recently, cross-lingual word embeddings have attracted a lot of attention, because they can capture semantic meaning of words across languages, which can be applied to cross-lingual tasks. Most methods learn a single mapping (e.g., a linear mapping) to transform word embeddings space from one language to another. In this paper, we propose an advanced method for improving bilingual word embeddings by adding a language-specific mapping. We focus on learning Japanese-English bilingual word embedding mapping by considering the specificity of Japanese language. On a benchmark data set of Japanese-English bilingual lexicon induction, the proposed method achieved competitive performance compared to the method using a single mapping, with better results being found on original Japanese words.**

***Keywords*-Cross-lingual word embeddings; Japanese; Word translation**

I. Introduction

Distributed representations of words, so-called word embeddings [1], [2], [3], have achieved impressive results in many Natural Language Processing (NLP) tasks and applications [4], [5], [6]. While it is possible to obtain monolingual word embeddings for most languages, the monolingual word embeddings in different languages are not comparable, hence, cannot be used in cross-lingual NLP applications.

With the needs of representing words in cross-lingual settings, several models for learning cross-lingual word embeddings have been proposed [7], [8], [9], [10]. These models can be divided mainly into two types: "online" and "offline". "Online" approaches learn word embeddings of two languages simultaneously by exploiting monolingual texts and some parallel corpora [11], [12]. "Offline" approaches firstly obtain the word embeddings in two languages independently, and then learn a mapping function (e.g., a linear mapping) from pre-trained monolingual word embeddings [7], [9], [13], [14], [15]. In this paper, we focus on this offline approach.

The early works of "offline" approaches learn bilingual word embedding mappings using some bilingual signals, such as bilingual lexicons [7], [9] and parallel or comparable data [8], [16], [17], as supervision to minimize distance between two word embedding spaces. These methods rely on bilingual lexicons or parallel corpora. To mitigate the needs of bilingual data, some recent methods focused on an unsupervised way, which attempts to learn cross-lingual word embeddings by mapping word embedding spaces to each other based on the distribution information of word embeddings [14], [18] or generative adversarial networks [19].

While both of supervised and unsupervised methods have shown impressive results, the shortcoming of these methods is that they have an assumption that entire word embeddings in one language are mapped to other language by following a same mapping function (e.g., a linear mapping). In this paper, we introduce a method to relax the above assumption by adding a language-specific mapping. Our method aims at learning more precise mapping across languages by considering the specificity of languages. Specifically, we focus on learning Japanese-English bilingual word embedding mapping by considering the specificity of Japanese.

The remainder of this paper is organized as follows. Section II introduces the previous work of learning bilingual word embedding mappings. Section III explains the proposed method. Section IV and Section V present the experimental settings and results analysis. Section VI concludes this paper and outlines future work.

II. Bilingual word embedding mapping

In this section, we first present notations that are used in this paper. Then, we introduce the previous methods of learning bilingual word embedding mapping using bilingual lexicons, which is the basis of our method for improving Japanese-English bilingual word embedding mapping.

A. Notation

Given pre-trained monolingual word embeddings of two languages, the goal of bilingual word embedding mapping is to learn a mapping that aligns the two pre-trained monolingual word embeddings. Let $x \in \mathbb{R}^d$, $y \in \mathbb{R}^d$ denote word vectors in the source and target language embedding spaces respectively. Let x_i and y_i denote word vectors of an actual word in the source and target language vocabularies respectively.

B. Linear mapping

Learning a linear mapping between two monolingual word embedding spaces was first introduced in [7]. They

978-1-7281-5014-7/19/$31.00

use a bilingual dictionary (usually most frequent words) of n = 5000 pairs of words with their associated vectors $\{x_i, y_i\}_{i\in\{1,n\}}$ to learn a linear mapping W between two monolingual word embeddings spaces by employing stochastic gradient descent to minimize the squared reconstruction error:

$$\min_{W} \sum_{i=1}^{n} \|y_i - Wx_i\|^2 \quad (1)$$

With the learned mapping W, any word embedding in the source language can be mapped to the target language embedding space by $y = Wx$.

Based on the linear mapping model that is introduced above, [9], [20] demonstrated that quality of bilingual word embeddings is improved by adding an orthogonality constraint on W in (1). Our method learns bilingual word embedding mapping based on the method in [9].

III. The Proposed method

In Section II, we introduced previous methods of learning a linear mapping between two language embedding spaces, which are based on an assumption that all word embeddings in one language follow the same mapping function. This single mapping can be seen as a global mapping, which is obviously a simplified method. While the global mapping performs good on the words that represent common concepts across languages (e.g., 犬 - dog), we noticed that the performance drops off for language-specific words (reported in Table II), such as some original Japanese words that their corresponding English words are their romanizations (e.g., 石田, a Japanese surname, its corresponding word in English is its romanization "ishida"). This shows the limitation of mapping all the word embeddings using a single global mapping, since global mapping does not consider the language specificity.

To solve the above problem, we propose an advanced method, in which we learn a language-specific mapping and combine it with global mapping. We believe that the combination of global mapping and language-specific mapping should achieve better performance.

In our method, we first learn a global mapping based on the method in [9], which is introduced in Section II. Then, we use the learned global mapping to extract a set of original Japanese words and their romanizations pairs (e.g., 石田 - ishida), which are used as a bilingual dictionary to learn a language-specific mapping. Finally, we transform word embeddings from one language to another by employing the combination of global mapping and language-specific mapping.

A. Global mapping

To learn the global mapping W_g between Japanese and English word embedding spaces, we follow the method in [9], which is a supervised method that needs some bilingual lexicons. While recent unsupervised methods achieve competitive or superior performance for some language pairs compared to supervised methods, their performance degrades for distant language pairs, such as Japanese-English, which was experimentally proved by [13], [15]. Thus, we choose a supervised method and use a set of Japanese words and their English translations as a bilingual dictionary to learn a global linear mapping between Japanese and English word embedding spaces.

B. Language-specific mapping

As introduced at the beginning of Section III, we are going to learn a language-specific mapping by employing language-specific words. We aim to utilize Japanese language-specific words, which are original Japanese words whose corresponding English words are their romanizations.

We extract original Japanese words in the following steps:

1) Identify original Japanese words candidates: All the Japanese words can be romanized and represented in Latin scripts. However, only the original Japanese words' corresponding English words should be their romanizations. Thus, we first roughly identify original Japanese words candidates by checking whether their romanizations appear in English vocabulary.

2) Identify Japanese origins of transliterated words in English: For some original Japanese words candidates that are obtained in the previous step, even their romanizations appear in English vocabulary, but their true corresponding English words might be not their romanizations. In order to filter out these words, we firstly utilize the global mapping to map all the candidates of original Japanese words to an English embedding space. Then, the mapped Japanese word vectors are compared with all the English word vectors to find the closest words in English. For a given Japanese word, if most of its closest words in English are transliterated words from Japanese words, it has more possibility to be an original Japanese word. Based on this observation, for each candidate of original Japanese word, we find its 5 closest words in English. Then, we identify transliterated words in these 5 closest English words by using the method in [21], which distinguishes a transliterated word from English words (non-transliterated words) by judging whether it can be segmented by Japanese syllabaries. In the experiments, for a given original Japanese word candidate, we empirically determine it as an original Japanese word if at least 3 of its 5 closest English words are transliterated words.

After this step, we can obtain a set of original Japanese words. We use the romanizations of these original Japanese words as their corresponding English words. In this way, we extract a set of pairs of original Japanese words and their romanizations, which are used as a bilingual dictionary to learn a language-specific mapping W_s based on the method in [9] that is introduced in Section II.

C. Combination of global mapping and language-specific mapping

For a given Japanese word, its word vector x is mapped to English word embedding space by using both global

mapping W_g and language-specific mapping W_s. When comparing x with a English word vector y, it obtains two similarity scores, which are represented as $S_{sim}(W_g x, y)$ and $S_{sim}(W_s x, y)$. The details of similarity calculation metrics we used are introduced in the implementation details in Section IV.

To combine the two similarity scores, we propose to either select the maximum score as the final similarity degree, or the weighted summation of two similarity scores:

$$S_{sim}(x, y) = \max\{S_{sim}(W_g x, y), S_{sim}(W_s x, y)\} \quad (2)$$

$$S_{sim}(x, y) = \alpha \cdot S_{sim}(W_g x, y) + (1 - \alpha) \cdot S_{sim}(W_s x, y) \quad (3)$$

$\alpha\,(\alpha \in [0, 1])$ is the weight that are used to balance the two similarity scores.

IV. Experiments

In this section, we experimentally evaluate our method in bilingual lexicon induction task, which measures the word translation accuracy in comparison to a gold standard.

A. Experimental dataset

We evaluated our method on the widely used MUSE dataset [19], which consists of dictionaries for many language pairs divided into training and test sets. We evaluated our method by inducing lexicons between Japanese and English. The training data set is composed of 5000 Japanese words with their English translations. The test data set is composed of 1500 Japanese words with their English translations.

The monolingual word embeddings were trained using *word2vec* with the skip-gram model [2]. The Japanese word embeddings were trained on Japanese Wikipedia corpus, and the English word embeddings were trained on English Wikipedia corpus.

B. Implementation details

1) Embedding normalization: Reference [9] proved that pre-processing of monolingual word embeddings with length normalization and dimension-wise mean centering can improve the performance of liner bilingual word embedding mapping. In our experiments, we follow the recommended settings in [9] to pre-process the monolingual word embeddings by applying firstly length normalization, then dimension-wise mean centering, and then length normalization again to ensure that the final embeddings have a unit length.

2) Retrieval method in bilingual lexicon induction: In bilingual lexicon induction task, given the words in the source language, it needs to have a retrieval metric that is used to select corresponding translations. Reference [22] demonstrated that nearest neighbors retrieval suffers from the hubness problem, which is that a few words (known as hubs) dominates as becoming nearest neighbors over many other words. Among the existing solutions to mitigate hubness [10], [19], we utilize the Cross-domain Similarity Local Scaling (CSLS) [19].

To measure the similarity between a mapped source word vector Wx and a word vector in the target language y, the CSLS considers the average similarity of Wx and y for their k nearest neighbors in another language, respectively. The CSLS similarity measure is defined as:

$$\begin{aligned}\text{CSLS}(Wx, y) &= 2\cos(Wx, y) \\ &- \frac{1}{k}\sum_{y' \in \mathcal{N}_{\text{T}}(Wx)} \cos\left(Wx, y'\right) - \frac{1}{k}\sum_{Wx' \in \mathcal{N}_{\text{S}}(y)} \cos\left(Wx', y\right)\end{aligned} \quad (4)$$

where $\mathcal{N}_{\text{T}}(Wx)$ is the set of k nearest neighbors of Wx from the target language. Similarly, $\mathcal{N}_{\text{S}}(y)$ is the set of k nearest neighbors of y. Following the parameter settings in [19], we set $k = 10$ in our experiments.

C. Baseline methods

We compare the proposed method with following baselines:

1) Global mapping (GM): This type of methods aims at transforming words in one language vector space to another by using a single global mapping. We use GM as a baseline to verify the effectiveness of our method, which combines a language-specific mapping with GM.

2) Language-specific mapping (LM): This method only use language-specific mapping to map Japanese word embedding to English. We compare our method with LM in order to examine the effect of only considering language specificity.

D. Our method

As introduced in the last subsection of Section III, our method combines similarity scores (CM) from global mapping and language-specific mapping in two ways: one (i.e., (2)) is to select the maximum score, which is represented as CM_{max}. The other (i.e., (3)) utilizes the weighted summation of two similarity scores, which is represented as CM_{sum}.

V. Results and analysis

A. Evaluation metrics

We evaluate the experimental results by using precision @1, @3, and @5. The precisions are equal to the rates of Japanese words whose correct corresponding English words are found in the top 1, 3, 5 results, respectively.

B. Overall experimental results

Table I shows the experimental results. For CM_{sum}, we show the results when $\alpha = 0.9$ as it has better performance than other weights. In addition, we report some sample results from the experiments of GM and CM_{sum} $(\alpha = 0.9)$ in Table III.

From Table I, we can that the overall performance of CM_{max} and CM_{sum} are on par with GM, which is the most competitive baseline. Comparing LM with other two methods, LM achives poor performance. It indicates that

Table I
PRECISION@N FOR JAPANESE-ENGLISH BILINGUAL LEXICON INDUCTION TASK ON MUSE DATASET COMPARED WITH BASELINE METHODS

	P@1 (%)	P@3 (%)	P@5 (%)
GM	36.38	48.08	52.58
LM	12.06	16.85	19.68
CM_{max}	36.75	48.37	52.87
CM_{sum} $(\alpha = 0.9)$	36.67	48.51	53.09

only using original Japanese words to learn a language-specific mapping is not enough for obtaining high-quality bilingual word embeddings.

C. Performance on original Japanese words

In order to evaluate the performance of our method on original Japanese words, we manually extract all the original Japanese words in the test data set. There are totally 499 word pairs that their Japanese words are original Japanese words.

Table II shows the experimental results over original Japanese words. CM_{sum} $(\alpha = 0.3)$ significantly outperforms the two baselines GM and LM. The statistical significance was confirmed in a two-sided t-test at a significance level of 0.1. Comparing the results of GM with other methods, the lowest performance is observed in GM over the original Japanese words, which indicates that using a single global mapping is a simplified method that cannot achieve good performance on language-specific words as discussed in the beginning of Section III.

D. Analysis of different weights for combination of GM and LM

In our method, when combining the similarity scores of global mapping and language-specific mapping, the weighted summation method CM_{sum} (i.e., (3)) needs to set the weight α to balance two similarity scores. To investigate how the weight α affects the performance of CM_{sum}, we conducted experiments on all test data set and original Japanese data set using different values of α.

Fig. 1 presents the results of using different weights to combine GM and LM. The higher values of α achieve better performance on the all test data set. However, the lower values of α obtain better performance on the data set of original Japanese words. This indicates that the impact of LM is more obvious on original Japanese words. In addition, this reveals the necessity of determining α based on data characteristics. For example, when data set contains more original Japanese words, the similarity score of LM should have a higher weight.

Table II
EXPERIMENTAL RESULTS ON THE DATA SET OF ORIGINAL JAPANESE WORDS

	P@1 (%)	P@3 (%)	P@5 (%)
GM	22.12	29.88	33.65
LM	29.18	36.94	40.71
CM_{max}	23.29	30.82	34.59
CM_{sum} $(\alpha = 0.3)$	34.12	44.00	48.24

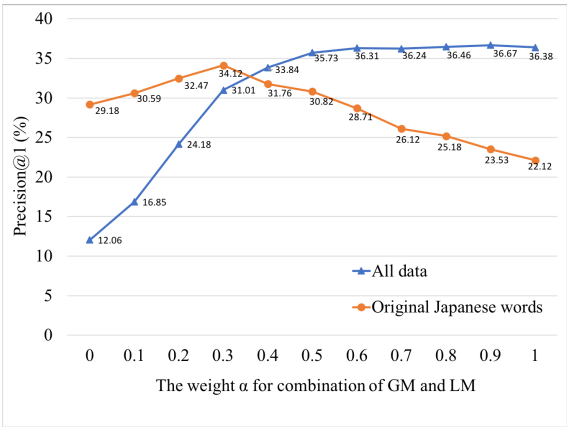

Figure 1. Experimental results of different weights for combining GM and LM

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a method to learn Japanese-English bilingual word embedding mapping. Our method exploits the specificity of Japanese to learn a language-specific mapping, which is combined with global mapping for improving Japanese-English bilingual word embedding mapping. The experimental results on the MUSE benchmark data set of bilingual lexicon induction task proved the effectiveness of the proposed method.

In the future, we will apply our method to other language pairs including Japanese and European languages. In addition, we plan to extend our method beyond the word level to longer text units such as phrases or sentences.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Number JP16K00452, and MEXT-Supported Program for the Strategic Research Foundation at Private Universities (S1511026).

REFERENCES

[1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[3] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[4] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning*, vol. 32, 2014, pp. 1188–1196.

[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 260–270.

Table III
SAMPLE RESULTS OF GM AND CM_{sum} $(\alpha = 0.9)$. "O" REPRESENTS ORIGINAL JAPANESE WORDS. BOLD TEXT INDICATES THE GROUND TRUTH, AND THE NUMBERS IN PARENTHESIS ARE THE RANKS OF THE GROUND TRUTH ENGLISH WORDS WITHIN THE RETURNED RESULTS.

Japanese word	**Ranked English words returned by GM**	**Ranked English words returned by CM_{sum} $(\alpha = 0.9)$**
知り合い	**acquaintance (1)**, reacquainted, befriended, friend, colleague, **acquaintances (6)**, friends, romantically	reacquainted, **acquaintance (2)**, befriended, friend, colleague, **acquaintances(6)**, romantically, acquainted
毎時	headways, **hourly (2)**, half-hourly, hour, off-peak, kmph, round-trips, daytimes	headways, **hourly (2)**, half-hourly, hour, off-peak, kmph, round-trips, daytimes
勝者	**winner (1)**, losers, victors, **winners (4)**, contenders, match-up, loser, loser's	**winner (1)**, losers, victors, **winners (4)**, match-up, contenders, loser, loser's
屋外	**outdoor (1)**, open-air, **outdoors (3)**, indoors, indoor, unheated, well-lit, floodlights	**outdoor (1)**, **outdoors (2)**, open-air, indoors, indoor, unheated, well-lit, tents
栄一 (O)	bronisław, tatsuo, isao, yasuo, kazuo, hellmut, andrzej, **eiichi (8)**	bronisław, isao, tatsuo, **eiichi (4)**, kazuo, yasuo, ichirō, hellmut
彦根 (O)	kawagoe, karatsu, kumamoto, **hikone (4)**, kurashiki, kakegawa, hirosaki, shizuoka	**hikone (1)**, karatsu, kawagoe, kumamoto, kakegawa, hirosaki, kurashiki, shizuoka
岸田 (O)	yukio, naoto, masahiko, **kishida (4)**, ichirō, fumio, shota, makiko	yukio, **kishida (2)**, naoto, masahiko, ichirō, fumio, makiko, yasushi
熱海 (O)	**atami (1)**, abashiri, otaru, karuizawa, hakone, kusatsu, enoshima, onsen	**atami (1)**, abashiri, karuizawa, otaru, hakone, kusatsu, onsen, enoshima

[6] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with Compositional Vector Grammars," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2013, pp. 455–465.

[7] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168v1*, 2013.

[8] I. Vulić and M. F. Moens, "Bilingual distributed word representations from document-aligned comparable data," *Journal of Artificial Intelligence Research*, vol. 55, pp. 953–994, 2016.

[9] M. Artetxe, G. Labaka, and E. Agirre, "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2289–2294.

[10] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," in *Proceedings of the 5th International Conference for Learning Representations (ICLR2017)*, 2017.

[11] S. C. AP, S. Lauly, H. Larochelle, M. M. Khapra, B. Ravindran, V. Raykar, and A. Saha, "An autoencoder approach to learning bilingual word representations," *Advances in Neural Information Processing Systems*, pp. 1853–1861, 2014.

[12] K. M. Hermann and P. Blunsom, "Multilingual models for compositional distributed semantics," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 58–68.

[13] A. Joulin, P. Bojanowski, T. Mikolov, H. Jegou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2979–2984.

[14] M. Artetxe, G. Labaka, and E. Agirre, "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*, 2018, pp. 1–10.

[15] C. Zhou, X. Ma, D. Wang, and G. Neubig, "Density matching for bilingual word embedding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 1588–1598.

[16] M. T. Luong, H. Pham, and C. D. Manning, "Bilingual word representations with monolingual quality in mind," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015, pp. 151–159.

[17] J. Guo, W. Che, D. Yarowsky, H. Wang, and T. Liu, "Cross-lingual dependency parsing based on distributed representations," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, pp. 1234–1244.

[18] H. Cao, T. Zhao, S. Zhang, and Y. Meng, "A distribution-based model to learn bilingual word embeddings," in *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, 2016, pp. 1818–1827.

[19] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in *Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018)*, 2018.

[20] C. Xing, D. Wang, C. Liu, and Y. Lin, "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*, 2015, pp. 1006–1011.

[21] Y. Song, B. Batjargal, and A. Maeda, "Recognition and transliteration of proper nouns in cross-language record linkage by constructing transliterated word pairs," *International Journal of Asian Language Processing*, vol. 27, no. 2, pp. 111–125, 2017.

[22] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," *arXiv:1412.6568*, pp. 1–10, 2015.