Exploiting Common Characters in Chinese and Japanese to Learn Cross-lingual Word Embeddings via Matrix Factorization

Jilei Wang [*†]
Tsinghua University, China
wangjileiruc@gmail.com

Shiying Luo [†]
Northeastern University, China
neulsy@hotmail.com

Weiyan Shi
University of California, Berkeley, USA
wyshi@berkeley.edu

Tao Dai
Tsinghua University, China
dait14@mails.tsinghua.edu.cn

Shu-Tao Xia [*]
Tsinghua University, China
xiast@sz.tsinghua.edu.cn

Abstract

Learning vector space representation of words (i.e., word embeddings) has recently attracted wide research interests, and has been extended to cross-lingual scenario. Currently most cross-lingual word embedding learning models are based on sentence alignment, which inevitably introduces much noise. In this paper, we show in Chinese and Japanese, the acquisition of semantic relation among words can benefit from the large number of common characters shared by both languages; inspired by this unique feature, we design a method named CJC targeting to generate cross-lingual context of words. We combine CJC with GloVe based on matrix factorization, and then propose an integrated model named CJ-Glo. Taking two sentence-aligned models and CJ-BOC (also exploits common characters but is based on CBOW) as baseline algorithms, we compare them with CJ-Glo on a series of NLP tasks including cross-lingual synonym, word analogy and sentence alignment. The result indicates CJ-Glo achieves the best performance among these methods, and is more stable in cross-lingual tasks; moreover, compared with CJ-BOC, CJ-Glo is less sensitive to the alteration of parameters.

1 Introduction

Word representation is critical to various NLP tasks, and the traditional one-hot representation, despite its simplicity, suffers from at least two aspects: the vector dimensionality increases with vocabulary size, leading to "curse of dimensionality"; more importantly, it fails to capture the semantic relation among words.

Due to the defects of one-hot representation, the majority of research interests now have switched to distributed word representation (also known as "word embedding"), which represents word as a real-valued vector. Represented as vectors, the semantics of words are better reflected, as the relatedness of words can be quantified using vector arithmetic.

To efficiently train word embeddings, a range of models have been proposed, most of them targeting to train monolingual word embedding. Though word embedding is often discussed under monolingual scenario, cross-lingual embedding can serve as a useful tool in several NLP tasks including machine translation (Wu et al., 2016), word sense disambiguation (Chen et al., 2014), and so on. This is because cross-lingual word embeddings map words from two languages into one vector space, thereby making it possible to measure the semantic relation among words from different languages. However, compared with the bulk of works studying monolingual word embedding, cross-lingual word embedding is still at its initial stage, with no learning model being widely accepted.

In this paper, we present a method named CJC (<u>C</u>hinese-<u>J</u>apanese <u>C</u>ommon Character)

[*] Corresponding authors.
[†] Contributed equally to the paper.

Proceedings of the 3rd Workshop on Representation Learning for NLP, pages 113–121
Melbourne, Australia, July 20, 2018. ©2018 Association for Computational Linguistics

aiming to extract cross-lingual context of words from sentence aligned Chinese-Japanese corpus. Given the large amount of common characters shared by both languages and the rich semantic connections thereof, we exploit them to acquire potential word level alignment. The acquired cross-lingual contexts can be flexibly integrated with various models; in this paper, CJC is mainly integrated with a matrix factorization model called Glove (Pennington et al., 2014), and the integrated model is thus called CJ-Glo.

To evaluate the performance of CJ-Glo, we take 2 sentence aligned models respectively based on CBOW(Mikolov et al., 2013a) and GloVe, and CJ-BOC model (based on Common Character + CBOW) (Wang et al., 2016) as contrast, and compare the trained word embeddings of these methods using three typical NLP tasks, including cross-lingual synonym, word analogy and sentence alignment. According to the experiment results, the acquired word embeddings by using CJ-Glo have better quality than those of the other models; moreover, CJ-Glo performs more stably than its competitors, and is less sensitive to parameter alteration.

2 Related work

Word embedding was initiated by Hinton (1986), which essentially encodes word using a real-valued vector. With word embeddings, the intrinsic relatedness among words can be explicitly measured as the distances or angles between word pairs. This favorable feature of word embedding soon led to its popularity in industry and academia in past decades. Specifically, word embedding has found its applications in machine translation (Wu et al., 2016; Lample et al., 2017), word sense disambiguation (Chen et al., 2014; Guo et al., 2014), information retrieval (Vulić and Moens, 2015) and so on.

To efficiently acquire high-quality word embeddings, vast research efforts have therefore emerged. A representative framework to learn word embeddings is Neural Network Language Model (NNLM) proposed by Bengio et al. (2003), which adopts back-propagation when training word embeddings and parameters for the model. Another typical approach is matrix factorization, whose basic idea is to approximate original matrices with low-rank matrices by leveraging statistic information. For example, GloVe (Pennington et al., 2014) explicitly factorizes the co-occurrence matrix, training only non-zero elements instead of an entire spare matrix.

Traditionally, word embedding was studied under monolingual setting, and then naturally extended to bilingual scenario. Compared with monolingual word embeddings, bilingual word embedding reveals the internal relation among words of different languages; and such capability makes bilingual word embeddings a powerful tool to assist machine translation, or even serves as a substitute for word mapping matrix and dictionary in previous machine translation methods. A range of works have been proposed to learn bilingual word embeddings, such as (Mikolov et al., 2013b), which attempts to map separately trained word embeddings into one vector space, and acquire bilingual word embeddings. BilBOWA is a model proposed in (Gouws et al., 2015), whose most notable merit is the whole training process does not require word alignment or dictionary. word alignment or dictionary. (Shi et al., 2015) is another work that utilizes matrix factorization in word embeddings learning. Ruder et al. (2017) provides a detailed survey, which enumerates the input format and basic principles of various bilingual word embedding learning methods.

When it comes to non-alphabet-based language like Chinese and Japanese, an essential difference from alphabet-based languages is that each character in a word contains abundant information, and makes sense itself. In addition to this, an underlying correlation between Chinese and Japanese is the large portion of shared characters in both languages; with the help of these characters, Chu et al. (2014) extracted texts from Wikipedia web pages of Chinese and Japanese version, based on which they then constructed a Chinese-Japanese parallel corpus. A natural conjecture about the common characters is the semantic similarity or even equivalence among them. In light of this, we proposed CJ-BOC model in our previous work (Wang et al., 2016) to learn Chinese-Japanese bilingual word embeddings, which outperforms sentence-alignment approaches in terms of embedding quality. To our knowledge, our previous work is the first attempt to learn Chinese-Japanese word embeddings using common Chinese characters.

3 Chinese-Japanese Common Character

Historically, Chinese character has spread to a group of countries in East Asia as a major carrier of Chinese culture, thereby influencing the writing systems in these countries. Traditional Chinese, Simplified Chinese and Japanese Kanji are now being used, all developing from Traditional Chinese; and given the same root of them, these three writing systems actually share a large portion of common characters: for a certain character in one of them, we can find its counterparts in the other two, with minor variation or even of the same shape. Chu et al. (2012) proposed a Chinese character table comparing traditional Chinese, simplified Chinese and Japanese. As summarized in Table 1, the glyphs of such common characters can be 1) the same in all these three writing systems; 2) consistent in two of them; 3) different in all these three.

And with regard to their semantics, simplified and traditional Chinese are only two written forms of the same language, and therefore common characters within them are semantically equivalent. For Japanese Kanji, most characters are semantically equivalent or relevant to their counterparts in Chinese.

We in our previous work (Wang et al., 2016) quantified such semantic relatedness from the view of information theory using mutual information (MI) and conditional mutual information (CMI). By repeating the experiments in this paper, we acquired the results in Table 2. All these 5 characters have multiple meanings in both Chinese and Japanese, and their respective meanings differ to some extent in both languages. Normally CMI should be larger than MI, which indicates that in a translation-sentence pair, if 2 words from each sentence share a common character, they are likely to form a translation word pair. The results of shown in Table 2 are no exception, providing theoretical root for our model which will be proposed in section 4.

4 Model

4.1 Context of Word and CJC Method

Before delving into the learning models, we should first clarify the concept of context. In natural language processing, a widely adopted semantic representation model is Bag-of-Words (Zhang et al., 2010). The fundamental assumption of this model is: within a given sentence or paragraph, the target word is prone to have the most intimate semantic relation with its closest context words. Formally define a sentence S with l words as an ordered sequence: $S = \langle w_0, w_1, ..., w_l \rangle$, and context function $Ctx(\cdot)$ is often formulated as:

$$Ctx(w_i, S) = \{w_k | i - K \leq k \leq i + K\}. \quad (1)$$

In cross-lingual scenario, besides two monolingual corpora of both languages, a parallel corpus is often required in most models, which is aligned in either word-level (Guo et al., 2016) or sentence-level. Some recent works attempted to learn embeddings without using parallel corpus, such as (Artetxe et al., 2017).

Now try to consider bilingual context of a given target word in aligned parallel corpus. Let $\langle S_{zh}, S_{ja} \rangle$ be a sentence pair, then define:

$$\begin{aligned} Ctx(w_{zh,i}) &= Ctx(w_{zh,i}, S_{zh}) \\ &\cup\ Ctx(w_{zh,i}, S_{ja}). \end{aligned} \quad (2)$$

As formulated above, the context of a target word is the union of its contexts in both sentences. Therefore in word-aligned parallel corpus, let $\langle w_{zh,i}, w_{ja,j} \rangle$ be a pair of aligned words, and the cross-lingual context $Ctx_w(w_{zh,i}, S_{ja})$ is equal to $Ctx_w(w_{ja,j}, S_{ja})$, since contexts in both languages are taken into account in this definition. In sentence-aligned parallel corpus, the cross-lingual context $Ctx_s(w_{zh,i}, S_{ja})$ is defined as the set of all the words in the respective sentence.

In real applications, sentence alignment data are usually easier to acquire. For example, Chu et al. (2014) proposed an approach to align Chinese-Japanese cross-lingual wiki corpus, using the common characters between both languages.

According to the analysis in Section 3, given an aligned Chinese-Japanese sentence pair, word alignment can be performed upon word pairs that share common characters. Based on

Type	Example of Characters with Unicode				Percentage
	SC	TC	KJ	Meaning	
1 - AAA	人 (U+4EBA)	人 (U+4EBA)	人 (U+4EBA)	People	56.55
2 - AAB	窗 (U+7A97)	窗 (U+7A97)	窓 (U+7A93)	Window	4.63
3 - ABA	国 (U+56FD)	國 (U+570B)	国 (U+56FD)	Country	3.45
4 - ABB	习 (U+4E60)	習 (U+7FD2)	習 (U+7FD2)	Study	29.17
5 - ABC	图 (U+56FE)	圖 (U+5716)	図 (U+56F3)	Picture	6.19

Table 1: Corresponding examples and percentages(%) of common characters in Simplified Chinese (SC), Traditional Chinese (TC), and Japanese Kanji (KJ).

Table 2: Estimated MI and CMI of 5 Common Characters.

	MI	CMI
天	0.3369	30.3057
地	0.5804	87.4515
人	0.8942	151.0069
中	0.7337	138.9676
学	0.4173	119.8921

this conclusion, using common characters, we can now give a definition for context similar to context in sentence-align corpus.

Define a character matching function $CC(\cdot)$ that generates a set of word in which each word has at least one common character with target word $w_{zh,i}$:

$$\begin{aligned} CC(w_{zh,i}, S_{ja}) = \; & \{w_{ja} | w_{ja} \in S_{ja}, \\ & c \in w_{zh,i}, c \in w_{ja}\}. \end{aligned} \tag{3}$$

Thus parallel context $Ctx_c(w_{zh,i}, S_{ja})$ can be acquired via common character matching:

$$\begin{aligned} Ctx_c(w_{zh,i}, S_{ja}) = \; & \{w | w \in Ctx(w_{ja}, S_{ja}), \\ & w_{ja} \in CC(w_{zh,i}, S_{ja})\}. \end{aligned} \tag{4}$$

Hence, when multiple words in the corresponding sentence have common characters with the target word, all of them will be included in $Ctx_c(w_{zh,i}, S_{ja})$. However, such case rarely occurs during our experiments.

For example, “天气/不错/一起/去/散步/吧” and “天気/が/良い/から/散歩/しまし/ょう” are a parallel sentence-pair, meaning “The weather is nice, let’s take a walk”. There are two corresponding word pairs detected by common characters: “天气-天気 (Weather)” and “散步-散歩 (Take a walk)”. Two words in a pair share their respective context during training.

We name this method as CJC (Chinese-Japanese Common Character) which uses $CC(\cdot)$ to determine context. Different from our previous work (Wang et al., 2016) which exploited common characters to facilitate only CBOW, this CJC method is more of a generalized scheme that can be integrated with various models including CBOW, Skip-Gram, GloVe etc.

4.2 CBOW-like Models

CBOW was a model proposed by Mikolov et al. in (Mikolov et al., 2013a), whose optimization goal is maximizing a probabilistic language model. In cross-lingual especially Chinese-Japanese scenario, the objective function for training $w_{zh,i}$ is:

$$\begin{aligned} L(S_{zh}) = \; & \tfrac{1}{N} \sum_{i=1}^{N} \Big\{ P_{zh,i,zh} \\ & + \; \lambda \cdot P_{zh,i,ja,c} \\ & + \; \mu \cdot P_{zh,i,ja,s} \Big\}, \end{aligned} \tag{5}$$

where $P_{zh,i,zh}$, $P_{zh,i,ja,c}$, $andP_{zh,i,ja,s}$ are softmax function of the target word $w_{zh,i}$ to its corresponding monolingual context, sentence aligned cross-lingual context, and CJC context. Both λ and μ here are parameters of the model. If $\lambda = 0$, this is a trial sentence aligned CBOW model, otherwise it is a CJC+CBOW model; the CJ-BOC model in our previous work (Wang et al., 2016) used similar approach, and would be used as a baseline in our experiments.

4.3 GloVe-like Models

4.3.1 GloVe

GloVe model was originally proposed by Pennington et al. (2014). As the name implies, GloVe utilizes the global information of the corpus for vector training. GloVe and CBOW, as commonly adopted learning models, however differ a lot in terms of mathematical models, as they are respectively based on matrix factorization and neural network. The process of GloVe is as follows:

First, construct a word-word cooccurence matrix $M = (m_{ij})_{n\times n}$, where n is the size of the corpus, and m_{ij} represents the number of occurrence of w_j in the context of w_i in all the sentences S.

The learning problem of GloVe can then be transformed into the optimization of function $F(\cdot)$, such that for any word embeddings x_i, x_j and probe word embedding $\widetilde{x}_k$, the objective function is defined below:

$$L=\sum_{i,j=1}^{n} f(m_{ij})(x_i^T\widetilde{x}_j+b_i+\widetilde{b}_j-\log m_{ij})^2, \quad (6)$$

$$f(m)=\begin{cases} (\frac{m}{m_{max}})^{\alpha} & \text{if } m<m_{max} \\ 1 & \text{otherwise.} \end{cases} \quad (7)$$

In this function, both b_i and b_j are bias, and f is a weighing function aiming to mitigate the impact of dataset size on training results. In GloVe, m_{max} is set to 100 and α to $\frac{3}{4}$.

4.3.2 Cross-lingual GloVe and CJ-Glo

To fit GloVe in cross-lingual scenario, one should first expand the word-word cooccurrence matrix. Suppose two languages respectively contain n and t words, the new matrix would have a size of If w_i and w_j belong to the same language, m_{ij} can be computed using exactly the same way as in GloVe; otherwise, suppose (S_{zh}, S_{ja}) is a pair of parallel sentences, $w_i \in S_{zh}$, $w_j \in S_{ja}$, and we have:

$$m_{ij}=\sum_{(S_{zh},S_{ja})}(\lambda\cdot C_{ij,c}+\mu\cdot C_{ij,s}), \quad (8)$$

$$\begin{aligned} C_{ij,c} &= Cnt(w_j, Ctx_c(w_i, S_{ja})), \\ C_{ij,s} &= Cnt(w_j, Ctx_s(w_i, S_{ja})), \end{aligned} \quad (9)$$

$Cnt(\cdot)$ counts the frequency of w_j in certain context of w_i, either sentence aligned context or CJC context.

Once the cross-lingual word-word cooccurence matrix is obtained, the following optimization unfolds similarly with the monolingual GloVe model, using the objective function (6) and weighting function (7) to train.

Similar to Cross-lingual CBOW model, if the CJC learning rate $\lambda = 0$ in equation (8), this is a sentence aligned cross-lingual GloVe model. Otherwise, it is a CJC-enhanced model, and is thus called CJ-Glo.

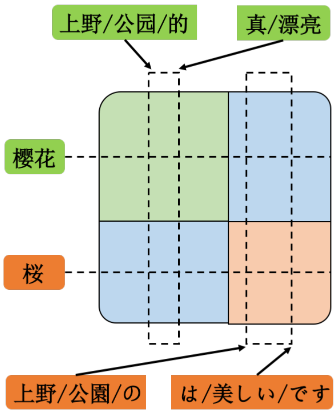

Figure 1: An example of CJ-Glo model, where the window size is 7 and the common character is "櫻 (桜)".

Figure 1 demonstrates the operational principle of CJ-Glo: the square in this figure is a cross-lingual word co-occurrence matrix, in which the green square is a Chinese monolingual co-occurrence sub-matrix, and the orange square is for Japanese. The blue sections are cross-lingual sub-matrices and elements in them are calculated using equation (8). When two parallel sentences each contain a word sharing common characters, each word would be taken as a co-occurrence in the context of the other. Every point crossed by dotted lines and dotted rectangles represents an element to increment when processing the sentence pair.

5 Experiments and Analysis

5.1 Evaluation Methods

To evaluate the quality of cross-lingual word embeddings obtained from various models, we conducted three groups of experiments: 1) the straightforward cross-lingual synonym comparison; 2) cross-lingual word analogy; 3) sentence alignment.

Cross-lingual synonym comparison.

In monolingual scenario, the word embeddings of a pair of synonyms should have a high cosine similarity. This property is also applicable in cross-lingual word embeddings, i.e., the cosine similarity between a word embedding and its translated counterpart should also be high. In real applications, the correspondence between words in source language and words in target language can be one-to-one, one-to-many, or vice versa. To effectively eliminate ambiguity, we picked 200 one-to-one corresponding word pairs $\langle w_{zh}, w_{ja} \rangle$ at random, then for each word pair, calculated the cosine similarity between w_{zh} and w_{ja}, denoted as d, and computed the rank of d among the cosine similarities from w_{zh} to every Japanese word in corpus V_{ja}. Use the rank to calculate its relative rate among all words:

$$rate = (1 - \frac{rank - 1}{total\_word\_num}) \times 100\%. \quad (10)$$

Conducted the same operation for w_{ja} and all words in corpus V_{zh}. Calculate the average rate for all the 200 word pairs, and acquire the average rate of $w_{zh} \rightarrow w_{ja}$ and $w_{ja} \rightarrow w_{zh}$ respectively. Ambiguity is eliminated in all these word pairs, so a large rate is therefore favored.

Cross-lingual word analogy.

Word analogy is probably the most widely adopted task to evaluate the performance of word embeddings, because it depicts the connection between trained vector space and word semantics. Both CBOW(Mikolov et al., 2013a) and GloVe(Pennington et al., 2014) used a dataset with 19,544 queries for evaluation.

Given several related words from different languages, cross-lingual analogical reasoning works as follows: y=v(はは)-v(ちち)+v(男孩), we hope that the relatedness between Japanese words "はは (mother)" and "ちち (father)" could help us find the Chinese word "女孩 (girl)" and Japanese "女の子 (girl)" through Chinese word "男孩 (boy)".

More formally, the cross-lingual analogy task was undertaken as follows:

1. Input a quadruple of word embeddings $\langle w_1 : w_2 :: w_3 : w_4 \rangle$, where each word could be either Chinese or Japanese;
2. Compute the target vector $u = w_2 - w_1 + w_3$, acquire the corresponding rank and rate as in cross-lingual synonym comparison for $u \rightarrow w_4$;
3. Based on the ratio of Chinese word count to Japanese word count in the quadruple $\langle w_1 : w_2 :: w_3 : w_4 \rangle$, the word analogy task is divided into 5 subtasks, whose ratio are $(0 : 4)$, $(1 : 3)$, $(2 : 2)$, $(3 : 1)$ and $(4 : 0)$, and their respective query amount is 420, 1680, 2520, 1680, and 420 in our experiment;
4. Calculate the average rate on every subtask.

Also, the average rate here is expected to be as large as possible.

Sentence alignment.

The above experiments respectively evaluated the direct similarity and cross-lingual feature of word embeddings. And now we consider a more complicated task: sentence alignment. In the dataset from (Chu et al., 2014), other than training data, a manual test dataset was also attached, which are 198 sentence pairs. Using this dataset, we conduct this experiment as follows:

1. For a Chinese sentence $S_{zh,i}$, calculate its average vector $U_{zh,i}$ and all U_{ja} of all sentences S_{ja}, and compute the cosine similarity.
2. Sort all the cosine similarities in step 2, and acquire the rank of the average vector $U_{ja,i}$ of $S_{ja,i}$ (the parallel sentence of $S_{zh,i}$).
3. Transform rank into rate using formula 10, where total number is 198.
4. Compute the average rate $S_{zh} \rightarrow S_{ja}$;
5. Follow the same steps above to generate $S_{ja} \rightarrow S_{zh}$.

Compared with the previous experiments, which evaluate only the relation between individual word embeddings, sentence alignment is a comprehensive task using word embedding, and is a critical indicator for the overall quality of the trained word embeddings.

5.2 Dataset and Training Details

As mentioned previously, (Chu et al., 2014) generated a parallel corpus including Chinese-Japanese sentence pairs from Wikipedia; train.ja and train.zh in this dataset were used throughout our empirical study, both containing 126,811 lines of text. Concretely, every single line in these two files is a complete sentence, which is parallel to its counterpart in the other file. As the preprocessing for datasets, both files were segmented using MeCab[1] and Jieba[2] for Japanese and Chinese, respectively. During the preprocessing, we assured the segmentation on Chinese and

[1] http://taku910.github.io/mecab, accessed date: December 20, 2017.

[2] https://github.com/fxsjy/jieba, commit number: cb0de2973b2fafaa67a0245a14206d8be70db515.

Table 3: Parameters of CJC and sentence learning rates in each models.

Model	λ	μ
SenBow	0	0.2
CJ-BOC	0.4	0.2
SenGlo	0	0.2
CJ-Glo	0.4	0.2

Table 4: Cross-lingual synonym comparison results on 200 one-to-one word pairs, the average rates(%) of each models.

Model	$w_{zh} \rightarrow w_{ja}$	$w_{ja} \rightarrow w_{zh}$
SenBow	83.97	83.76
CJ-BOC	96.75	97.61
SenGlo	91.17	90.05
CJ-Glo	**97.97**	**98.80**

Japanese were approximately grained, by tuning parameters.

Four models in total are put into comparison in our experiment:

1. *SenBow* model is the bilingual CBOW model applying sentence-aligned method;

2. *CJ-BOC* model from (Wang et al., 2016), considered as a CJC+CBOW model;

3. *SenGlo* model applies sentence-aligned method to GloVe model;

4. *CJ-Glo* model is our CJC method enhanced GloVe model.

The parameters of CJC learning rate λ and sentence learning rate μ are showed in Table 3. Both SenGlo and CJ-Glo have a m_{max} of 100, and an α of $\frac{3}{4}$. The thread count is 16 in the implementations of all these four models, the output vector dimensionality is 100, and the training process is iterated 15 times. We set the parameters to the above values, since these models achieved the optimal performances under such settings in our evaluation. All models are implemented using C language, and the code can be found on GitHub[3].

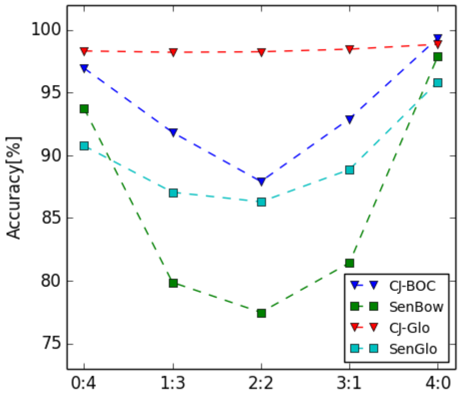

Figure 2: Cross-lingual word analogy experiment result. X-axis is the number ratio of Chinese words and Japanese words in the analogy query ($w_1 : w_2 :: w_3 : w_4$).

5.3 Results

The result of cross-lingual synonym comparison is shown in Table 4, from which we can see the integration of Common Character leads to obvious performance improvement for both CBOW-like and GloVe-like models, compared with sentence-aligned models, and CJ-Glo achieve the best result.

Figure 2 summarizes the results of the cross-lingual word analogy task, whose X-axis represents the ratio of Chinese word count to Japanese word count. In the figure, the leftmost point represents the result of pure Japanese word analogy, and the rightmost is the pure Chinese word analogy. We can see that all 4 models achieve fair performances in pure Chinese/Japanese word analogy. However, when it comes to the cross-lingual word analogy, CJ- models outperform Sen- models, and GloVe-like models generally beat CBOW-like ones. Another noticeable fact is that CJ-Glo performs approximately good under all 5 ratios, showing basically no difference between cross-lingual and monolingual word analogy.

We display the sentence alignment results in Table 5. Similarly, we still find CJ- models outperform Sen-, and GloVe-like models beat CBOW-like ones. Again, CJ-Glo has the best performance.

According to the above experiments, we can see compared with typical sentence-aligned methods, Common Character enhanced mod-

[3] https://github.com/jileiwang/CJC, commit number: a10592d200bc15f7b53d81a8f895e7de9ef8676d.

Table 5: Sentence alignment results on 198 parallel sentence pairs, the average rates(%) of each models.

Model	$S_{zh} \rightarrow S_{ja}$	$S_{ja} \rightarrow S_{zh}$
SenBow	79.14	74.63
CJ-BOC	86.39	83.14
SenGlo	90.33	84.90
CJ-Glo	**91.57**	**86.00**

els are superior in learning Chinese-Japanese cross-lingual word embeddings, as it achieves obvious performance boost in various tasks. Moreover, CJ-Glo performs better than CJ-BOC, and is non-sensitive in cross-lingual tasks.

5.4 Model Analysis: CJC Learning Rate

CJC learning rate here refers to the multiplying factor of CJC context $Ctx_c(\cdot)$, which is λ in CJ-BOC and CJ-Glo. It worths discussion that how would CJC learning rate affects the performance of our proposed models. To explore this issue, we conduct a simple experiment: fixing the other parameters as set in section 5.2, we only change CJC learning rate, and apply the acquired word embeddings to synonym $w_{zh} \rightarrow w_{ja}$ tasks. The results are displayed in Figure 3 , in which we can find as λ increases in CJ-BOC, the accuracy declines after an increase, showing a obvious local optimal. While in CJ-Glo, the accuracy keeps improving with the increase of λ. Note that both parameters should be less than 1, because otherwise the impact of cross-lingual context would dominate the learning process, obviously resulting in overfit. CJ-Glo is more stable during the change of CJC learning rate, this interesting difference between both models is related to the their underlying learning mechanisms.

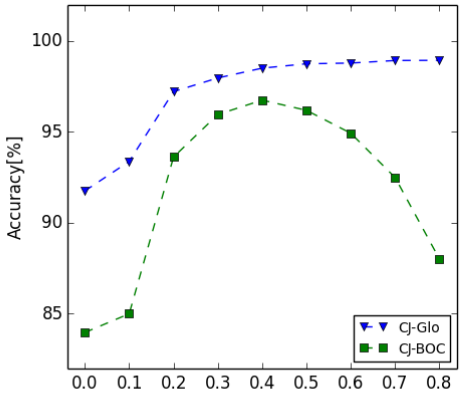

Figure 3: Accuracy of CJ-BOC and CJ-Glo Models on cross-lingual synonym $w_{zh} \rightarrow w_{ja}$ with different CC learning rate.

6 Conclusion and Future Work

In this paper, we quantified the semantic connection among common characters shared by Chinese and Japanese, and utilized it as the theoretical root to propose our cross-lingual context extracting method CJC. CJC makes use of common characters of both languages to assist the acquisition of parallel contexts. The effectiveness of CJC enhanced matrix factorization model CJ-Glo was verified via a series of tasks including cross-lingual synonym, word analogy and sentence alignment. As the experiment result shows, models like CBOW and GloVe achieved notable performance gain after integrated with CJC. Furthermore, CJ-Glo performed the best among all evaluated state-of-the-art methods, and showed its stability on cross-lingual tasks and non-sensitiveness of training parameter changing.

Below are several directions we may work on in the future: 1) The idea of training character and word embeddings jointly (Chen et al., 2015) is applicable to Chinese-Japanese word embedding training. Meanwhile, we can also align common characters and train cross-lingual character embeddings to further improve the quality of trained word embeddings. 2) A recent work (Lai et al., 2016) indicates that the performances of a model may vary given different tasks. Therefore, we shall study the performance fluctuation of CJ-Glo with more tasks including machine translation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant Nos. 61771273, 61371078.

The authors would like to thank Xinyu Zhou, Minghao Li, Yanning Li, and Qi Tang for the insightful discussions and suggestions that helped us improve the manuscript.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 451–462.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1025–1035.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In *Proceedings of the 24th International Joint Conference On Artificial Intelligence (IJCAI)*, pages 1236–1242.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2012. Chinese characters mapping table of japanese, traditional chinese and simplified chinese. In *Proceedings of the 8th Conference on International Language Resources and Evaluation Conference (LREC)*, pages 2149–2152. Citeseer.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Constructing a chinese japanese parallel corpus from wikipedia. In *Proceedings of the 9th Conference on International Language Resources and Evaluation Conference (LREC)*, pages 642–647.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 748–756.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING)*, pages 497–507.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *AAAI*, pages 2734–2740.

Geoffrey E Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA.

Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. 2016. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 567–572.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372. ACM.

Jilei Wang, Shiying Luo, Yanning Li, and Shu-Tao Xia. 2016. Learning chinese-japanese bilingual word embedding by using common characters. In *International Conference on Knowledge Science, Engineering and Management*, pages 82–93. Springer.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52.