Machine Translation (2021) 35:503–518
https://doi.org/10.1007/s10590-021-09274-0

Improving bilingual word embeddings mapping with monolingual context information

Shaolin Zhu[2] · Chenggang Mi[1] · Tianqi Li[2] · Fuhua Zhang[2] · Zhifeng Zhang[2] · Yu Sun[2]

Received: 25 February 2020 / Accepted: 16 June 2021 / Published online: 21 July 2021

Abstract
Bilingual word embeddings (BWEs) play a very important role in many natural language processing (NLP) tasks, especially cross-lingual tasks such as machine translation (MT) and cross-language information retrieval. Most existing methods to train BWEs are based on bilingual supervision. However, bilingual resources are not available for many low-resource language pairs. Although some studies addressed this issue with unsupervised methods, monolingual contextual data are not used to improve the performance of low-resource BWEs. To address these issues, we propose an unsupervised method to improve BWEs using optimized monolingual context information without any parallel corpora. In particular, we first build a bilingual word embeddings mapping model between two languages by aligning monolingual word embedding spaces based on unsupervised adversarial training. To further improve the performance of these mappings, we use monolingual context information to optimize them during the course. Experimental results show that our method outperforms other baseline systems significantly, including results for four low-resource language pairs.

Keywords Bilingual word embeddings · Low-resource · Unsupervised emthod

1 Introduction

Bilingual word embedding (BWE), which aims to find word translations of different languages, has recently become a major focus of NLP research. BWEs can be learned from two pre-trained monolingual word embeddings via supervised or unsupervised methods Mikolov et al. (2013b); Smith et al. (2017); Zhang et al.

✉ Chenggang Mi
michenggang@nwpu.edu.cn

[1] Northwestern Polytechnical University, Xi'an 710129, China

[2] Zhengzhou University of Light Industry, Zhengzhou 450002, China

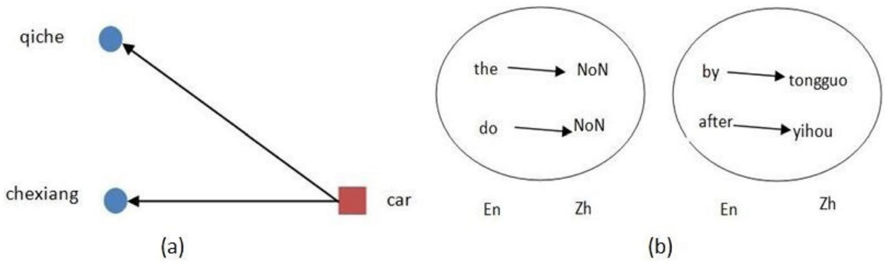

Fig. 1 An illustration of bilingual words for translating from English to Chinese

(2017a, b); Conneau et al. (2018); Patra et al. (2019). These studies have shown the advantages of bilingual word mapping in word translation induction for many low-resource language pairs.

There are two steps in traditional BWE model training: (1) Training word embeddings for each language. (2) Mapping word embedding of two languages into a shared space. Two groups of methods that can map one word embedding to another: supervised and unsupervised. Mikolov et al. (2013b) first proposed that word embeddings exhibit similar structures across languages. They learned a linear mapping from a source to a target embedding space by employing a parallel vocabulary of five thousand words as anchor points. Recent attempts at reducing the need for bilingual supervision Smith et al. (2017); Zhang et al. (2017a), Patra et al. (2019) to train bilingual word mappings use as few bilingual resources as possible have been shown to reach a comparable performance in several cross-lingual NLP tasks.

Although many researchers have shown the effectiveness of their methods on connecting two monolingual word embeddings during BWE model training, most of them require bilingual data as supervision, either in the form of parallel corpus or seed lexicon. Unfortunately, bilingual resources are not available for many low-resource language pairs. Hence, recent proposed unsupervised methods explore distribution-based approaches Cao et al. (2016) or adversarial training Zhang et al. (2017b) to obtain cross-lingual word embeddings without any bilingual data. These studies rely on a basic assumption that the words in different languages should have a similar distribution Mikolov et al. (2013b). Despite they have a similar distribution, Søgaard et al. (2018) showed that these spaces are, in general, far from being isomorphic. Conneau et al. (2018) used frequent words to refine the results of adversarial learning via Procrustes analysis. Then, they extracted a synthetic dictionary from the shared embedding space by a new method called cross-domain similarity local scaling(CSLS). Although it is an effective strategy to obtain word translations, two issues still exist: (1) For example, for a common lingual phenomenon: synonyms problem. The monolingual multiple synonyms have the same translation in another language (see Fig. 1a). (2) That only using frequent words as anchor points to refine mappings may introduce new noise. Many functional English frequent words do not actually have translations in Chinese, but there are also some words that have a translation (see Fig. 1b).

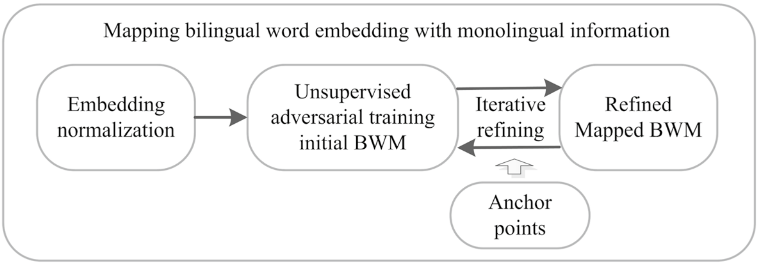
Mapping bilingual word embedding with monolingual information
Embedding normalization
Unsupervised adversarial training initial BWM
Iterative refining
Refined Mapped BWM
Anchor points

Fig. 2 A general architecture of our model

To sum up, unsupervised methods still have some flaws in mapping bilingual word embeddings. In fact, the hubness issue can be effectively relieved by mining monolingual context Zhang et al. (2016); Patra et al. (2019). Pointwise mutual information(PMI) can effectively find the internal relationship between words such as synonyms Khan et al. (2016). We propose to utilize PMI to obtain a more monolingual context to refine bilingual word embeddings.

In this paper, we propose to improve the mappings of bilingual word embeddings without employing any bilingual resources. This method is very useful for low-resource language pairs. We first train monolingual word embeddings separately on two large monolingual corpora. Motivated by the GAN Goodfellow et al. (2014), we then formulate our task by adversarial training to learn embedding mappings from two monolingual corpora, one in the source and one in the target language. The adversarial training process contains a generator and discriminator. The generator tries to learn transformed distribution from source language embeddings and makes the distribution lie close to the embedding of target language. The discriminator is a binary classifier that strives to distinguish between the transformed distribution and the target embedding. After that, we use the Procrustes analysis to refine the initial distribution to get a better bilingual word embedding. Ren et al. (2014) used Procrustes analysis to align images and got a significant improvement in image detection. In this paper, we can apply the Procrustes analysis to refine bilingual word embeddings. However, how to select suitable anchor points heavily affects Procrustes analysis to align two embedding spaces. We propose two strategies PMI Khan et al. (2016) and TD-IDF to select anchor points to refine bilingual mapping via Procrustes analysis. During training, our approach alternates between refining bilingual word embeddings and selecting the anchor points accordingly (see Fig. 2).

Finally, our evaluation of the bilingual lexicon induction task reveals significant performance compared to all baselines on real-world datasets. We show that our strategies substantially improve bilingual word embedding. This achievement in turn allows our method to succeed, which is particularly favorable for low-resource language pairs.

In summary, this paper makes the following main contributions:

(1) We propose a new unsupervised method that maps bilingual word embeddings without parallel data. Our method tries to utilize monolingual context effectively

to obtain a hopefully better bilingual mapping. In addition, we demonstrate the effectiveness of our approach by experiments on four low-resource language pairs where parallel corpora are not available.

(2) We introduce a criterion to select more suitable anchors to refine bilingual word embeddings. This approach can significantly improve the bilingual mapping, especially in low-resource language pairs.

The paper is organized as follows. Section 2 describes the details that carry out embedding normalization, adversarial training and how to select suitable anchor points to optimize bilingual word embedding. We then present our training settings in Sect. 3. We report in Sect. 4 our results on bilingual lexicon induction tasks for several language pairs and compare our approach to baselines. Finally, we explain how our approach differs from recent related work on learning bilingual word embeddings.

2 Methodology

In this section, we introduce our method in detail. Our goal is to map bilingual word embeddings from two large scales of monolingual corpora without supervision. We first train monolingual word embeddings with fastText Bojanowski et al. (2017) separately on monolingual corpora. Then, adversarial training maps an initial bilingual word embedding space. Finally, we iteratively refine the bilingual space by mining the monolingual context.

Formally, we use X_s and Y_t to denote the word embedding of source and target language, respectively. Our goal is to learn the linear transformation matrice W_X so that the source embedding $W_X X_t$ lies close to the embedding of the target Y_t. The learned linear transformation matrices can be used to denote that two words in different languages have similar semantics. As illustrated in Fig. 2, our method contains three parts:

(1) *Embedding normalization*. Our method starts with preprocessing that length normalizes the embeddings, mean centers each dimension, and then orthogonalization revises each embedding.

(2) *Adversarial training initial BWE*. An initial BWM embedding is trained with adversarial training. This step is the basis of unsupervised refining of bilingual word mappings. We minimize source and target embedding space as follow:

$$Loss = argmin\|W_X X_s - Y_t\| \tag{1}$$

(3) *Refining BWE*. We expect a better BWE embedding than the initial state. In order to reach this goal, we adjust local bilingual mapping by optimizing and selecting anchor points.

2.1 Embedding normalization

Embedding normalization has been shown effective in previous work Artetxe et al. (2016), while Smith et al. (2017) showed that imposing an orthogonal constraint on the linear operator leads to better performance. Using embedding normalization has several advantages. First, It ensures that the dot product of any two embeddings is equivalent to their cosine similarity and directly related to their Euclidean distance. Then, a more stable model can be trained by normalizing embedding. In training word embedding, we use a similar method as Conneau et al. (2018). The rule on the embedding W:

$$W = (1 + \alpha)W - \alpha(WW^T)W \tag{2}$$

According to Conneau et al. (2018), $\alpha = 0.01$ is a suitable value to perform the next step. Then, we transform embedding into a normalized orthogonal basis. We use Schimidt normalization Jagadeesha et al. (1994) to do this, the transformation can be defined as follow:

$$W_n \leftarrow W_n - \frac{[W_1, W_n]}{[W_1, W_1]}W_1 - \dots - \frac{[W_{n-1}, W_n]}{[W_{n-1}, W_{n-1}]}W_{n-1} \tag{3}$$

2.2 Bilingual word embeddings initialization

The underlying difficulty of the mapping between two word embedding matrices X_s and Y_t is unaligned across both spaces without parallel corpora. As the embeddings are trained respectively on two languages, two embeddings matrices X_s and Y_t are difficult to align in one space directly. In order to overcome this challenge and build an initial bilingual embedding space, we use a generative adversarial network to map two monolingual word embeddings into one space Goodfellow et al. (2014); Zhang et al. (2017a).

From the above section, two sets of monolingual word embeddings can be got. Next, our goal is learning an initial mapping that connects two embeddings into one space, the mapping matrix W_X makes $Y_t \approx W_X X_s$. In this paper, we employ an adversarial training to implement the mapping. The adversarial training contains a generator and a discriminator. This approach is in line with the work of Conneau et al. (2018), which proposed to learn latent representations invariant to the input language. The generator G tries to make a mapping matrix to confuse the discriminator. In order to get orthogonal parametrization, we note that transforming the source word embedding into the target, its transpose should also transform the target to the source. The $D(source = 1|x)$ means that a representation x can be mapping into a source embedding. The generator can be formulated by minimization as follow:

$$L_G = -logD(source = 0|W_X X_s) - logD(source = 1|Y_t) \tag{4}$$

The discriminator D is a binary classifier which aims to enhance its ability to distinguish Y_t and $W_X X_s$. The discriminator can be implemented by maximizing as follows:

$$L_G = -logD(source = 1|W_X X_s) - logD(source = 0|Y_t) \tag{5}$$

In fact, the similar distribution assumption is correct roughly so that the above adversarial training procedure only captures some cross-lingual mapping. In our English-Chinese experiments, the average cosine similarity is better than a random solution. While the result is far from being useful on its own (the accuracy of the resulting dictionary is only 41.8%, as Table 1), it is substantially better than chance, and it works well as an initial solution for the refining method described next.

2.3 Refining bilingual word embeddings

An initial bilingual word embedding can be learnt by Sect. 2.2. However, only using the initiation, we can't induce a positive bilingual lexicon from the bilingual embedding spaces (as experimental results shown in Table 1). Procrustes analysis is applied to align two shapes, Ren et al. (2014) used it to implement face alignment. In this paper, we exploit the Procrustes analysis to refine our bilingual mapping. It mainly contains two steps. First, we detect the feature vectors of two embedding spaces as anchor points; Second, we use the Procrustes analysis to transfer one embedding space to another to align them.

In detail, we first assume that we have already detected feature vectors as anchor points. Next, for source embedding matrix p_s and target matrix q_t, our goal is that rotate, scale and translate p_s to coincide p_s and q_t. We use s, t, R to denote scaling, translation, and rotation, respectively. This problem can be formalized as a *Loss*:

$$Loss = \left\| sRp_s^T + t - q_t^T \right\|_F \tag{6}$$

Where p_s^T is the transpose of p_s, q_t^T is similar. The above formulation can be minimized as follow:

$$\underset{s,R,t}{argmin} \left\| sRp_s^T + t - q_t^T \right\|_F \tag{7}$$

$\|.\|_F$ is a Frobenius norm, which is the sum of squares. At the same time, Smith et al. (2017) show that an orthogonal constraint can get more stable results. Therefore, we add an orthogonal constraint as follow:

$$R^T R = I \tag{8}$$

I is an identity matrix. Ren et al. (2014) has proven that the above process can obtain an optimal solution.

We mainly introduce how to use Procrustes analysis to align the two embedding spaces described in the above section. However, we ignore how to select suitable anchor points and we will focus on anchor points in the rest of this section. Conneau et al. (2018) proposed to use frequent words as anchor points to implement alignment via Procrustes analysis. Their experimental results show that selecting frequent words as anchor points outperforms the adversarial approach. However, this method has two drawbacks as described in Sect. 2.1. So we can't expect to achieve promising

Table 1 Accuracy (%) of the proposed method in comparison with previous work

Methods	en-fr	en-es	en-zh	en-tr	zh-uy	zh-ka	zh-ne	zh-si	en-ne	en-si
NoN-Adv(Mikolov et al.)	25.6	23.2	17.3	14.2	10.21	9.52	8.86	9.15	11.7	12.1
NoN-Adv(Artetxe et al.)	32.7	31.6	18.5	16.3	10.68	9.8	9.63	10.12	12.35	12.82
Adv(Zhang et al.)	66.8	65.2	41.8	31.1	21.56	20.75	20.36	21.03	24.6	24.86
Adv-Refine-CSLS(Coneau et al.)	73.6	72.3	56.6	41.32	36.73	33.47	32.57	31.82	37.16	37.72
Proposed method	74.1	73.2	58.6	43.8	39.13	36.62	36.08	35.93	40.53	40.96

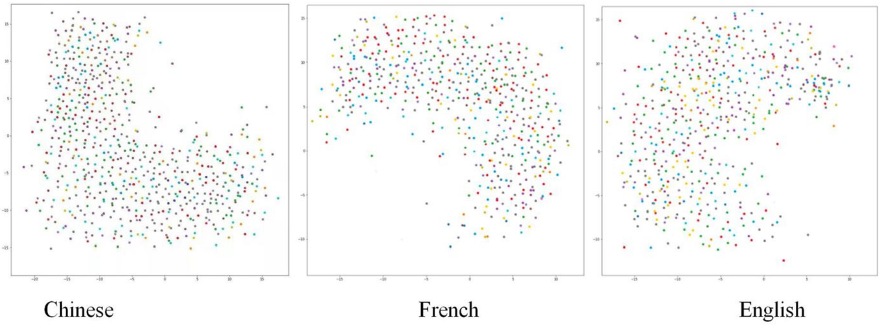
Chinese French English

Fig. 3 An illustration of monolingual and bilingual word embeddings on English-Chinese, **a** two monolingual embeddings map into a low-dimensional space. **b** Adversarial training bilingual embedding map into a low-dimensional space

performance in low-resource settings. In this paper, we propose two strategies mainly focus on filtering unsuitable anchor points.

Our first strategy is using the term frequency-inverse document frequency (TF-IDF) to filter anchor points. As we are known, every language has some special function words that are mainly used to connect, enhance and standardize semantic, such as some stop words in English or ”de” ”le” ”ba” in Chinese. Although some languages have listed some special function words, there is no such list in low-resource language pairs. TF-IDF is a good idea in information retrieval to identify or remove such function words. We count the frequency of words in sentences. TF-IDF can be formularized as follow:

$$TD-IDF = log\frac{N}{N(W_i)}.\left(log\frac{N+1}{(N(W_i)+1}+1\right) \tag{9}$$

N is the total number of words in the corpus, W_i is the number of i-th word. In experiments, we set the value of TF-IDF as 0.5 according to experience (Fig. 3).

In high-dimensional spaces, some vectors, dubbed hubs, are with high probability nearest neighbors of many other points, while others (anti-hubs) are not nearest neighbors of any point (as Fig. 4). This so-called hubness problem will affect to align two spaces. Most traditional approaches hinge on cross-lingual signals to link independent monolingual spaces: each word is associated with a vector that comprises monolingual statistics like pointwise mutual information (PMI), then the monolingual vector spaces are connected through bilingual signals. PMI is an important factor that measures the correlation between two variables, such as two words. We use PMI to mitigate the hubness problem in selecting anchor points. We can filter some points that tend to be nearest neighbors of many points. For two words w_i and w_j, we use PMI to calculate their correlation:

$$PMI(w_i, w_j) = log\left(\frac{P(w_i|w_j)}{P(w_j)}\right) \tag{10}$$

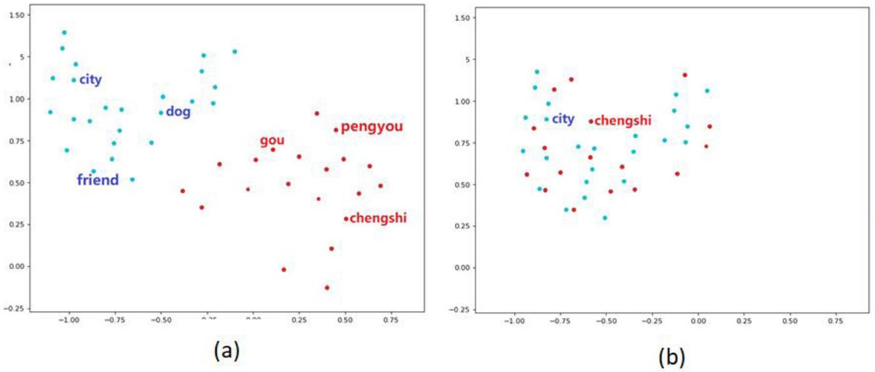

Fig. 4 An illustration of monolingual and bilingual word embeddings on English-Chinese, **a** two monolingual embeddings map into a low-dimensional space. **b** Adversarial training bilingual embedding map into a low-dimensional space

$P(w_j)$ is a probability of w_j and $P(w_i|w_j)$ is a conditional probability. In order to clarify the TF-IDF and PMI how to filter anchor words, we use an example in the next. For Chinese and English et al. they all often contain some stop words. They usually have a high word frequency and are useless for aligning word embeddings. We first use TF-IDF to filter those stop words. Then, we use PMI to solve the hubness problem. In detail, the word ”car” of English can be translated into ”qiche” and ”chexiang” of Chinese. However, this one-to-many may lead to the wrong alignment. The two words ”qiche” and ”chexiang” in Chinese have similar semantics. So we use PMI to filter those words that have similar semantics. In this paper, we mainly select some words with a large semantic gap as anchor words.

3 Experiments setting

In this paper, we carry out experiments in the widely used dataset from Wikipedia[1] monolingual corpora on four language pairs: English-Spanish, English-French, English-Chinese and English-Turkish. In addition, we test our model on six low-resource language pairs[2] (English-Nepali, English-Sinhala, Chinese-Nepali, Chinese-Sinhala, Chinese-Uyghur and Chinese-Kazakh). Following Conneau et al. (2018), we retain only words that occur at least 6000 times in our corpora. For Chinese, we use OpenCC[3] to normalize characters to be simplified, and then perform Chinese

[1] https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/.

[2] ftp://ftpmirror.your.org/pub/wikimedia/dumps/newiki/.

[3] https://github.com/BYVoid/OpenCC.

word segmentation with Jieba[4]. The preprocessing of English and French involves tokenization and lower casing which we carry out with the NLTK[5] toolkit. For Turkish, we utilize the preprocessing tools (tokenization and POS tagging) provided in LORELEI Language Packs (Strassel and Tracey 2016). For Nepali, Sinhala, Uyghur and Kazakh, as those languages are lack for preprocessing tools, we only remove punctuations.

Bilingual lexicon induction (BLI) is the most popular evaluation task for BWE that be used by previous work. As Faruqui and Dyer (2014); Cisse et al. (2017) showed that the nearest neighbor suffers from the hubness problem, so we adopt the Cross-domain Similarity Local Scaling (*CSLS*) from Conneau et al. (2018). Given two mapped embeddings x and y, the idea of *CSLS* is to compute $avg(x)$ and $avg(y)$, the average cosine similarity of x and y for their k nearest neighbors in the other language, respectively. Following the authors, we set $k = 10$. We can get a bilingual lexicon by using *CSLS*. In order to evaluate the bilingual lexicon, we need gold a standard for reference. The evaluation of ours is as follows: first, we use Google Translate to translate the source side vocabulary; second, the translations in the target language are queried again in the reverse direction to translate back to the source language, and those that don't match with the original source words are discarded.

We perform comparison experiments with previous studies: Zhang et al. (2017b) and Conneau et al. (2018). In the case of Conneau et al. (2018), we test the default hyperparameters in the source code as well as those reported in the paper, with iterative refinements. Given that Zhang et al. (2017b) report using a different value of their hyperparameters for different language pairs, we perform 10 runs for each, and report the best accuracies.

4 Results and discussion

In this section, we analysis experimental results and give some discussions.

4.1 Overall performance

As proof of evaluation on bilingual word embedding is BLI, we train our models and baselines using monolingual corpora. In order to verify our method can tackle the limitation of low-resource language pairs, we add supervised methods Mikolov et al. (2013b); Artetxe et al. (2016). We implement experiments according to their paper and use an empty list as a seed lexicon. Table 1 reports the results of bilingual lexicon induction in different methods.

From Table 1, we can observe that the two supervised methods get very poor performance when we set an empty seed lexicon. Although Mikolov et al. (2013b) and Artetxe et al. (2016) reported that they obtain a stable and good result, they depend on thousands of bilingual words or some special bilingual data so that those

[4] https://pypi.org/project/jieba/.

[5] http://www.nltk.org.

approaches are not suitable for low-resource language pairs. For rich-resource language pairs, there are enough bilingual data as supervision to improve the bilingual word embeddings. However, low-resource language pairs often lack bilingual data. This means that current supervised methods may don't obtain a good result for low-resource language pairs(we also can observe this point from Table 1). In all language pairs, our method outperforms the two supervised methods in the low-resource situation, which demonstrates the effectiveness of our method on low-resource language pairs.

Next, we report the results of Zhang et al. (2017b). on monolingual data (Zhang et al. 2017b carried out their experiment on comparable corpora). As it can be seen, our proposed method gets substantially better results on all language pairs than baselines. Then, we compare our method with Conneau et al. (2018). The reason is that we all refine the results of adversarial training. Although the accuracy improves on all language pairs, the improvements are various in different language pairs. Although we only got 0.5% and 0.9% on English-French and English-Spanish. We observe a more significant improvement in English-Chinese and English-Turkish, the value is 2% and 2.48%. The reason is that we use monolingual rather than comparable corpora, two embedding spaces have a more serious dissimilar distribution problem. Different languages make some points tend to be the nearest neighbors of many points in respective high-dimensional spaces Faruqui and Dyer (2014). We also give the word distributions of different languages in Fig. 3. We can find that English-French has a more similar distribution than English-Chinese. As English-Chinese is distant language pair, this means that the two languages are etymologically distant. The differences of distribution affect the bilingual word embedding mappings. When the two distributions are not similar, the similar assumption is not suitable for distant language pairs.

Moreover, we carry out the experiments in four realistic low-resource language pairs to test the effectiveness of our method. The performances of our approach are 39.13%, 37.62%, 38.53% and 38.96% on Chinese-Uyghur, Chinese-Kazakh, English-Nepali and English-Sinhala, which are 2.4%, 3.15%, 3.37% and 3.24% better than the best baseline. Although our method can obtain better results than baselines in realistic low-resource language pairs, we find the overall accuracy still is very low. Rich-resource language pairs have higher accuracies than low-resource language pairs. We count the numbers of training data of different languages(as in Table 2). We can find that low-resource languages have fewer articles than rich-resource languages. The size of training data is very important for the performance of BWE Zhang et al. (2017b). We then compare the performance of ours with Conneau et al. (2018), the main difference is the anchor points selection. Conneau et al. used frequent words as anchor points to refine bilingual embedding, we add some strategies to optimize the selection of anchor points. We explain that our method first tries to align the global distributions and then focuses on specific areas that may cause the hubness problem. In the next section, we will carry out two experiments to verify our hypothesis.

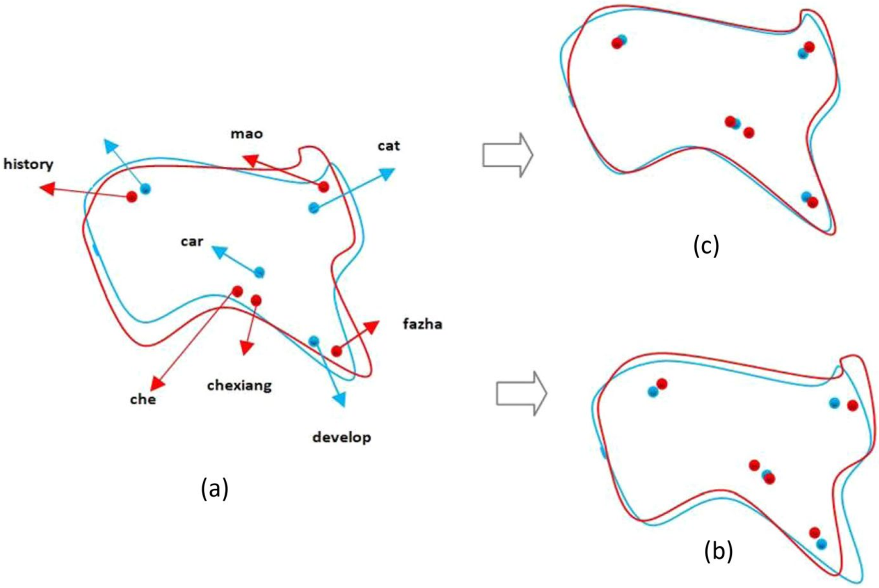

Fig. 5 An illustration of selecting different anchor points to refine bilingual embeddings

4.2 Word distribution

In order to investigate the necessity of refining the result of adversarial training, we run adversarial training to map two monolingual embeddings into a bilingual space. We record the two monolingual word distributions and adversarial bilingual word distributions, shown in Fig. 4. The plot is for English-Chinese, other language pairs exhibit similar results.

The adversarial training for mapping bilingual embedding without refining aligning bilingual space does not present a significant aligned space. Purely unsupervised methods, on the one hand, suffer from poor performance if the distribution of embedding spaces of two languages is very different from each other. Moreover, unsupervised methods can successfully align clusters of words, but miss out on fine grained alignment within the clusters (that so-called hubness problem). As we carry out experiments on monolingual corpora rather than comparable corpora (Zhang et al. 2017b used comparable corpora to carry out adversarial network), two monolingual embeddings are consistently able to align two spaces between the large blue and red clusters but have trouble aligning the smaller embedding sub-clusters.

4.3 Impact of anchor points

In this section, we investigate how anchor points selection results affect the performance of the refined bilingual embedding mappings. We refine bilingual embedding space via using frequent words and our method to select anchor points. Figure 5

Table 2 Training set statistics

	en	fr	es	zh	tr	uy	ka	ne	si
unzipped file size(GB)	26	11	8.1	4.1	1.5	1.3	1.2	1.4	1.4
number of articles(K)	5,690	2,301	1,560	1,619	397	316	308	363	356
number of words(K)	67	65	46	120	32	52	51	42	45

shows the different alignment for English-Chinese, other language pairs exhibit similar results.

Figure 5a is the alignment of adversarial training, Fig. 5b is the refined alignment via frequent words. We use our strategies to select anchor points and get Fig. 5c. It can be seen that the two methods all get a better performance compared with adversarial training. However, different anchor points have various alignments. In our method, we use TF-IDF and PMI to filter some useless or noisy anchor points so that the alignment is better. In Fig. 5, ”chexiang” and ”che” are common words and counted as frequent words in Chinese, ”Car” is a common English word, ”chexiang” and ”che” can be translated into ”Car”. As we select frequent words that exceed 6000 times as anchor points rather than the number of words, it makes the number of words of respective languages very disproportionate. Similar semantics are distinguished by different words in Chinese, while one word often combines with collocations to express different semantics. This phenomenon makes many Chinese words cluster, which in turn affects refining mappings (as Fig. 5a). We also count the number of words in different languages (As Table 2). We observe that although the size of English and French corpora exceeds Chinese, the number of words is much less. The statistical results also present why the improvement is more obvious than English-French and English-Chinese. For the low-resource language pairs, the hubness problem is very obvious due to data sparsity. Our method mitigates the hubness problem by PMI to filter some high similarity words when we select anchors to refine bilingual mappings. Therefore, our method obtains better performance on low-resource language pairs.

5 Related work

Many researchers have put their efforts into bilingual word embeddings mapping. Mikolov et al. (2013b) first used 5000 words as anchor points to learn this mapping and evaluated their approach on a word translation task. They also revealed the fact that these spaces exhibit similar distribution across languages. Since then, several approaches have been proposed to optimize the bilingual word embedding mappings Faruqui and Dyer (2014); Xing et al. (2015); Ammar et al. (2016); Zhang et al. (2017b). Other works on this topic attempted to use a few seeds lexicon to achieve the mapping Vulic and Korhonen (2016); Smith et al. (2017). Although those approaches can get a significant performance, Most of them rely on bilingual resources to provide supervision which is unavailable in low-resource language pairs.

Recently, unsupervised methods that map bilingual word embeddings were trained without using any manually created bilingual resources Cao et al. (2016); Barone (2016); Smith et al. (2017); Conneau et al. (2018). Cao et al. (2016) didn't require cross-lingual data to train bilingual word embeddings. They mapped word embeddings by matching the mean and variance of the Gaussian distribution. Smith et al. (2017) employed identical character strings to eliminate the bilingual limitation. Although these methods show encouraging results in practice, they also have a strong assumption on the writing systems of languages(e.g. that they need a common alphabet or Arabic numerals). Artetxe et al. (2018) proposed a self-learning algorithm that iteratively improves the mappings. Moreover, those methods rely on the assumption that the two embeddings are isomorphic. However, Søgaard et al. (2018) showed that these spaces are, in general, far from being isomorphic, although they have a similar distribution.

A few recent works attempt adversarial training for cross-lingual embedding transformation. Zhang et al. (2017b) employed adversarial training to map bilingual word embedding. Their approach relies on sharp drops of the discriminator accuracy for model selection. Conneau et al. (2018) adopted similar adversarial training to map bilingual word embedding. However, they used multiple ways repeat to refine the mapping, such as refining the mapping with the closed-form Procrustes analysis. Our model also employs adversarial training to initialize bilingual word embeddings. Then, we propose two strategies to filter anchor points and use Procrustes analysis to refine the initialization.

6 Conclusion

In this paper, we propose an unsupervised method to improve bilingual word embedding. Our method leverages adversarial training to learn an initial bilingual word embedding without any cross-lingual resources. Next, we make full use of monolingual context to select more suitable anchor points, then refine bilingual word embedding. In the experiments, we show that our model significantly and consistently outperforms baselines. Experimental results have shown the effectiveness of using the method on low-resource language pairs.

In the future, we plan to explore the following directions:

(1) As word distribution affects mapping bilingual word embedding, we will explore how domain classification affects bilingual word embedding.
(2) In this paper, we built a bilingual lexicon using our bilingual word embedding. We will experiment on other NLP and MT tasks to further research on its impacts.

Acknowledgements This work is supported by the National Natural Science Foundation of China (61906158), the Project of Science and Technology Research in Henan Province (212102210075).

References

Ammar W, Mulcaire G, Tsvetkov Y, Lample G, Dyer C, Smith NA (2016) Massively multilingual word embeddings, arXiv preprint arXiv:1602.01925

Artetxe M, Labaka G, Agirre E (2016) Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp 2289–2294

Artetxe M, Labaka G, Agirre E (2018) A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. arXiv preprint arXiv:1805.06297

Barone AVM (2016) Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. arXiv preprint arXiv:1608.02996

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146

Cao H, Zhao T, Zhang S, Meng Y (2016) A distribution-based model to learn bilingual word embeddings. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp 1818–1827

Cisse M, Bojanowski P, Grave E, Dauphin Y, Usunier N (2017) Parseval networks: Improving robustness to adversarial examples. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp 854–863

Conneau A, Lample G, Ranzato M, Denoyer L, Jégou H (2018) Word translation without parallel data. arXiv preprint arXiv:1710.04087

Faruqui M, Dyer C (2014) Improving vector space word representations using multilingual correlation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp 462–471

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680

Jagadeesha S, Sinha S, Mehra D (1994) A recursive modified gram-schmidt algorithm based adaptive beamformer. Signal process 39(1–2):69–78

Khan FH, Qamar U, Bashir S (2016) Sentimi: Introducing point-wise mutual information with sentiwordnet to improve sentiment polarity detection. Appl Soft Comput 39:140–153

Mikolov T, Le QV, Sutskever I (2013b) xploiting similarities among languages for machine translatio. arXiv preprint arXiv:1309.4168

Patra B, Moniz JRA, Garg S, Gormley MR, Neubig G (2019) Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. arXiv preprint arXiv:1908.06625

Ren S, Cao X, Wei Y, Sun J (2014) Face alignment at 3000 fps via regressing local binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1685–1692

Smith SL, Turban DH, Hamblin S, Hammerla NY (2017) Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859

Strassel S, Tracey J (2016) Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp 3273–3280

Søgaard A, Ruder S, Vulić I (2018) On the limitations of unsupervised bilingual dictionary induction. arXiv preprint arXiv:1805.03620

Vulic I, Korhonen A-L (2016) On the role of seed lexicons in learning bilingual word embeddings

Xing C, Wang D, Liu C, Lin Y (2015) Normalized word embedding and orthogonal transform for bilingual word translation. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1006–1011

Zhang M, Liu Y, Luan H, Sun M, Izuha T, Hao J (2016) Building earth mover's distance on bilingual word embeddings for machine translation. In: Thirtieth AAAI Conference on Artificial Intelligence

Zhang M, Liu Y, Luan H, Sun M (2017b) Adversarial training for unsupervised bilingual lexicon induction. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol 1. (Long Papers), pp 1959–1970

Zhang M, Peng H, Liu Y, Luan H, Sun M (2017a) Bilingual lexicon induction from non-parallel data with minimal supervision. In: Thirty-First AAAI Conference on Artificial Intelligence

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.