# Cross-lingual Feature Extraction from Monolingual Corpora for Low-resource Unsupervised Bilingual Lexicon Induction

**Zihao Feng**[†], **Hailong Cao**[†,*] **Tiejun Zhao**[†], **Weixuan Wang**[‡], **Wei Peng**[‡]
[†]Faculty of Computing, Harbin Institute of Technology
[‡]Artificial Intelligence Application Research Center, Huawei Technologies Co., Ltd
fengzihaogl@outlook.com
{caohailong, tjzhao}@hit.edu.cn
{wangweixuan2, peng.wei1}@huawei.com

## Abstract

Despite their progress in high-resource language settings, unsupervised bilingual lexicon induction (UBLI) models often fail on corpora with low-resource distant language pairs due to insufficient initialization. In this work, we propose a cross-lingual feature extraction (CFE) method to learn the cross-lingual features from monolingual corpora for low-resource UBLI, enabling representations of words with the same meaning leveraged by the initialization step. By integrating cross-lingual representations with pre-trained word embeddings in a fully unsupervised initialization on UBLI, the proposed method outperforms existing state-of-the-art methods on low-resource language pairs (EN-VI, EN-TH, EN-ZH, EN-JA). The ablation study also proves that the learned cross-lingual features can enhance the representational ability and robustness of the existing embedding model.

## 1 Introduction

Bilingual Lexicon Induction (BLI) has aroused great interest in the NLP research frontier. BLI aims to induce word translation pairs by aligning word embeddings trained independently from monolingual corpora. BLI has contributed to many NLP tasks, including unsupervised machine translation (Artetxe et al., 2018b), cross-lingual dependency parsing (Guo et al., 2015) and cross-lingual information retrieval.

Unsupervised BLI has achieved reasonable results compared with semi-supervised works in high-resource language settings, in which adversarial training were used in Lample et al. (2018); Zhang et al. (2017). These methods focused their attention on every single word, thus ignoring the relevance between words. Artetxe et al. (2018a) proposed a method (VecMap) using a similarity matrix as an initial solution to learn the second-order structural similarity of the embeddings. Another study directly leveraged an aligned similarity matrix instead of using the embedding matrix (Alvarez-Melis and Jaakkola, 2018). Recently, Peng et al. (2021) proposed a robust refinement technique based on the $\ell_1$ norm training objective. The methods above learned bilingual spaces by orthogonally projecting one monolingual space to another. Yet, evidence suggests that monolingual spaces, especially those of etymologically and typologically distant languages, are far from isomorphic (Søgaard et al., 2018; Vulić et al., 2019; Patra et al., 2019). Glavaš and Vulić (2020) relaxed the orthogonality constraint to improve the performance of BLI further. Mohiuddin et al. (2020) depicted a non-linear method using an encoder and decoder to learn the mapping in the latent space. Wang et al. (2019) proposed a joint training method using word alignments from parallel corpora as the supervision signals to align multilingual contextualized representations. While the methods mentioned above can leverage pre-trained embeddings in BLI, they lack the means to incorporate richer information like cross-lingual features from monolingual data. When it comes to low-resource and non-cognate language, the characterization capability of pre-trained embeddings is limited, which leads to the degeneration of these models.

On another strand of work, traditional works for BLI used the statistical methods to search the cross-lingual signals (Rapp, 1999; Koehn and Knight, 2002; Fung and Cheung, 2004; Gaussier et al., 2004; Haghighi et al., 2008; Vulić et al., 2011; Vulić and Moens, 2013). In recent years, E and Zhou (2022) proposed a more robust method. They formally defined the semantic embedding of words in a mathematical way instead of machine learning. However, their method is limited to non-topic words and requires high quality monolingual data to achieve good results.

The methods mentioned above have been reliant

[*]Corresponding author

*Proceedings of the 29th International Conference on Computational Linguistics*, pages 5278–5287
October 12–17, 2022.

on a high-resource language condition. Regarding low-resource and non-cognate language pairs, the characterization capability of pre-trained embeddings is limited, as only short-distance dependencies are available. In this paper, we propose a novel unsupervised method for BLI based on cross-lingual feature extraction. Furthermore, we design two ways to integrate the cross-lingual features with the pre-trained embeddings, which show complementary effects according to the experimental results. In summary, this paper makes the following contributions:

- We propose a method (CFE) to extract cross-lingual feature of each word (In Section 3.1). We expect the words with the same meaning in different languages have the similar representations and we can use it to initialize the UBLI directly.

- We propose two combination methods, embedding combination (ECB) and similarity combination (SCB), to use cross-lingual feature and pre-trained embeddings together to initialize the UBLI (In Section 3.2). The ECB method concatenates two kinds of embeddings by row, the SCB method weights the second-order similarity of pre-trained embeddings and the first-order similarity of the cross-lingual feature.

- Extensive experiments show that our method exceeds all previous unsupervised and state-of-the-art approaches on low-resource and distant language pairs. The ablation study shows that our cross-lingual feature is complementary to pre-trained embeddings. Our method improves the representational ability of the existing model (In Section 5).

## 2 Background

In this section, we describe the basic formulation of related supervised and unsupervised BLI methods. Let $X, Y \in \mathbb{R}^{n\times d}$ represent word embedding matrices in two languages $L_1$ and $L_2$, where $n$ is the number of words and $d$ is the dimension of the word embedding.

The key to supervised BLI is the parallel lexicon between two languages. Let $X^*, Y^* \in \mathbb{R}^{k\times d}$ represent parallel embedding matrices, say $x_i^*$ in $X^*$ is translated to $y_i^*$ in $Y^*$. Mikolov et al. (2013) pointed out that a linear transformation $W^*$ could be used to map two monolingual embeddings to a shared space.

$$W^* = \underset{W\in\mathbb{R}^{d\times d}}{\arg\min} \|X^*W - Y^*\|_F^2 \tag{1}$$

Artetxe et al. (2016) solved Problem (1) by adding an orthogonal constraint on $W$. Therefore, there is a closed-form solution to this problem called Procrutes: $W = UV^\top$, where $U$ and $V$ are defined by the SVD decomposition of $Y^\top X$.

For unsupervised BLI, embedding matrices $X$ and $Y$ are totally out of order. Therefore, unsupervised BLI needs a permutation matrix $P \in \mathcal{P}_n = \{0,1\}^{n\times n}$ to shuffles the row of $Y$:

$$\min_{W\in\mathbb{R}^{d\times d}, P\in\mathcal{P}_n} \|XW - PY\|_F^2 \tag{2}$$

Problem (2) can be solved by minimizing $W$ and $P$ in an iterative way. Grave et al. (2019) proposed a stochastic algorithm to initialize $W$ and $P$ randomly and estimate them in a joint way. However, effectively minimizing $P$ is hard. The key to unsupervised BLI is how to solve $P$ approximately.

Lample et al. (2018) proposed an adversarial method to initialize the initial dictionary. Artetxe et al. (2018a) has shown that two equivalent words in different languages have a similar distribution. Therefore, they initialized matrix $P$ based on similarity matrices of monolingual embeddings $M_X = XX^\top$ and $M_Y = YY^\top$. They used the second-order similarity of pre-trained embeddings to obtain a better initial dictionary. Then they applied a self-learning strategy to iteratively compute the optimal mapping and retrieve bilingual dictionary until convergence. Very recently, Wang et al. (2019) proposed a method to jointly train word embeddings on concatenated corpora of different languages and achieved good results.

In summary, the foundation for BLI is the parallel lexicon for initialization, especially for unsupervised BLI. Therefore, a high-quality initialization is the key for unsupervised BLI.

## 3 Methodology

In this section, we propose a novel framework to solve the problems of UBLI in low-resource scenarios. First and foremost, the CFE is used to extract cross-lingual representations from monolingual data. Then the ECB and the SCB are used to integrate the cross-lingual features with pre-trained embeddings, either by concatenation or similarity weighting.

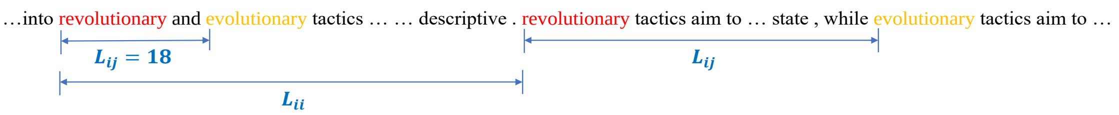
…into revolutionary and evolutionary tactics … … descriptive . revolutionary tactics aim to … state , while evolutionary tactics aim to …
$L_{ij} = 18$
$L_{ij}$
$L_{ii}$

Figure 1: Counting character level distance between word patterns. In this case, Let $x_i$ be the word "revolutionary" and $x_j$ be the word "evolutionary". For the first $L_{ij}$ in this example, we give the distance 18. Specifically, the distance of it is added up with 13 (the length of word "revolutionary"), 3 (the length of word "and") and 2 (the length of two space). The other $L_{ij}$ and $L_{ii}$ is calculated in the same way.

### 3.1 Cross-lingual Feature Extraction (CFE)

We propose a method to extract semantic information from monolingual data in this section. We think that although words have different symbolic representations in different languages, they all have the same language-independent textual features, which we call semantic feature. Empirically, when we are reading a novel, we understand a word based on the contextual information we can remember. So we define the semantic feature of a word based on the semantic relevance between it and its contextual words. Numerically, the semantic relevance between words are based on the character distance, which means distance counted in characters between words in a sentence. Let $x_i, x_j \in X$ represent two words in monolingual data. We define $L_{ij}$ as the character distance between the first letter of the word $x_i$ and $x_j$. When the word $x_i$ is the same as the word $x_j$, the distance is defined as $L_{ii}$. Note that we only calculate the closest $x_j$ after $x_i$. Figure 1 shows an example of $L_{ij}$. We count $L_{ij}$ for each word pair in monolingual data and define $n_{ij}$ as the number of $L_{ij}$ (We only count those $L_{ij}$ less than a certain threshold).

We find that semantic relevance is sensitive over short distances but degrades over long distances. For instance, when the word $x_i$ and $x_j$ is in different paragraphs, their semantic relevance is weak for all distances. So, we dropout some $L_{ij}$ through a threshold and define $S_{ij}$ to represent the semantic relevance between the word $x_i$ and $x_j$:

$$S_{ij} = e^{-\frac{L_{ij}}{D}} \tag{3}$$

Where $D$ is the hyperparameter. For every word pair, the number of their appearance also influences semantic relevance. Therefore, we weigh every $S_{ij}$ by $n_{ij}$:

$$\langle S_{ij} \rangle = \frac{n_{ij} \times S_{ij}}{\sum_{k=1}^{n} n_{ik} \times S_{ik}} \tag{4}$$

Here, $\langle S_{ij} \rangle$ denotes the average of $S_{ij}$ weighted by $n_{ij}$. Through this, we expect the $\langle S_{ij} \rangle$ represents strong semantic relevance. At last, for each word $x_i$, we extract $k$ words with the maximum semantic relevance in set $\langle S_{i*} \rangle \cup \langle S_{*i} \rangle$, where $\langle S_{i*} \rangle$ represents all word pairs start with word $x_i$, and $\langle S_{*i} \rangle$ is in the same way. In this way, we get the cross-lingual feature (The representations of words with same meaning in different languages are similar) of each word in $X$ and $Y$ separately through monolingual data:

$$Sem\_vec_{xi} = (\langle S_{i1} \rangle, \langle S_{i2} \rangle, ..., \langle S_{ik} \rangle) \tag{5}$$

Equation (5) denotes the cross-lingual feature of a specific word $x_i$, where $k$ is the hyperparameter, denotes the dimension of semantic vector.

### 3.2 Unsupervised initialization

Previous works VecMap (Artetxe et al., 2018a) has shown the effect of high-quality word pairs on unsupervised BLI. In this method, Let $X_{sem}, Y_{sem} \in \mathbb{R}^{n \times k}$ denote the cross-lingual feature matrices extracted by using the method in Section 3.1. We combine the cross-lingual feature to initialize unsupervised BLI in two ways, Embedding combination (ECB) and Similarity combination (SCB):

**Embedding combination (ECB):** We combine the pre-trained embedding with the cross-lingual feature as the initial embedding:

$$\begin{aligned} X_{com} &= X | X_{sem} \\ Y_{com} &= Y | Y_{sem} \end{aligned} \tag{6}$$

Where | denotes concatenation by row, therefore $X_{com}, Y_{com} \in \mathbb{R}^{n \times (d+k)}$. We follow the method in VecMap to calculate second-order similarity matrix as the initialization. In order to maintain the process of vector alignment is comparable with the other works. We only use $X_{com}$ and $Y_{com}$ to do the initialization step. We continually use $X$ and $Y$ to do the iterative process.

**Similarity combination (SCB):** For our feature $X_{sem}$ and $Y_{sem}$ is cross-lingual, we consider to calculate the similarity of $X_{sem}$ and $Y_{sem}$ directly. We combine this similarity with second-order similarity of $X$ and $Y$:

$$\lambda * [(XX^{\top})(YY^{\top})^{\top}] + (1-\lambda) * (X_{sem}Y_{sem}^{\top}) \quad (7)$$

Where $\lambda$ controls the ratio of two kinds of similarities. In this method, we can fully exploit the advantages of both kinds of embeddings.

Our method guarantees to converge to a loacl optimum base on the initial dictionary, so the quality of it is the key factor for our method. However, simply concatenating two kinds of embeddings and searching the nearest neighbor normally did not work in our preliminary experiments. It cannot guarantee to avoid our method getting stuck in poor local optima. For this reason, we propose some key improvements to make our initialization robust:

- **Embedding normalization:** The embeddings we use to be combined are trained in different ways. So, their meanings are completely different. When we combine them in a simple concatenation way, the representation ability of each embedding will be weakened. For this reason, we use a linear method ($\frac{X-Mean(X)}{Max(X)-Min(X)}$) to map two kinds of embeddings to $[-1, 1]$, then the combination method is more significant than before.

- **CSLS distance:** To extract the lexicon, we use the nearest neighbor for every word to search transformed embeddings. This phenomenon is known to occur the hubness problem (where one word is the nearest to many words) (Radovanovic et al., 2010; Suzuki et al., 2013). To avoid this hubness problem, Lample et al. (2018) modified it with the Cross-domain Similarity Local Scaling (CSLS). For two aligned embeddings $x$ and $y$, they denote the set $\mathcal{N}_T(Wx)$ and $\mathcal{N}_S(y)$ of the embeddings' k nearest neighbors in the other language, respectively. Then compute $r_T(x)$ and $r_S(y)$, the average cosine similarity of $\mathcal{N}_T(Wx)$ and $\mathcal{N}_S(y)$. The CSLS score of $x$ and $y$ can be computed as $CSLS(x, y) = 2cos(x, y) - r_T(x) - r_S(y)$. Following the authors, we set $k = 10$.

A high-level overview of our proposed method is outlined in Algorithm 1.

**Algorithm 1:** CFE method for UBLI

```
Input: monolingual corpora $L_1$ and $L_2$
Output: parallel dictionary
$X, Y \leftarrow$ pre-trained embeddings of $L_1, L_2$;
$X_{sem}, Y_{sem} \leftarrow$ cross-lingual feature extracted from $L_1, L_2$;
$Sim \leftarrow$ similarity matrix from ECB and SCB methods;
$D \leftarrow$ initial word translation dictionary using $Sim$;
while not convergence do
    $W \leftarrow$ linear mapping matrix calculated by Procrustes on $D$;
    $D \leftarrow$ CSLS($WX, Y$);
end
```

## 4 Experimental settings

In this section, we first list the baselines we used in Section 4.1, then we show the details of our own dataset and compare them with the MUSE dataset in Section 4.2. Finally, we show the hyperparameter settings of our methods in Section 4.3.

### 4.1 Baselines

We take several representative works of unsupervised BLI as our baselines. We choose the methods using pre-trained embeddings (Lample et al., 2018; Artetxe et al., 2018a; Li et al., 2020) and statistical method (E and Zhou, 2022) to be compared with our method. Especially, the method (Li et al., 2020) is the state-of-the-art model in low-resource languages and Peng et al. (2021) achieves a good results on high-resource language pairs using $\ell_1$ norm optimisation on refinement. Compared with all the baselines, our method uses both two kinds of features to initialize the BLI problem. We evaluate all the baselines by using MUSE parallel data and CSLS distance to do the nearest neighbor search. We execute the publicly accessible code or reproduce the code on our own to acquire the baseline findings due to the use of our own dataset.

| | EN | VI | TH | ZH | JA |
|---|---|---|---|---|---|
| words | 2418 | 326 | 33 | 288 | 298 |
| sentences | 72494 | 10677 | 273 | 1908 | 9110 |

Table 1: Details of the dataset. We show the number (K) of words and sentences in the monolingual corpus. For the words, we count the number of different words.

| Dataset | EN-VI | | EN-TH | | EN-ZH | | EN-JA | |
|---|---|---|---|---|---|---|---|---|
| | → | ← | → | ← | → | ← | → | ← |
| MUSE | 0.73 | 0.73 | 0 | 0.08 | 0.08 | 0 | 1.03 | 32.67 |
| Our Dataset | 0 | 0 | 0.11 | 0 | 0.08 | 0.28 | 43.80 | 31.64 |

Table 2: Results of VecMap on MUSE dataset (Lample et al., 2018) and our own dataset 4.2. We perform 10 runs for each experiment and report the average score of the accuracies (%).

| Model | EN-VI | | EN-TH | | EN-ZH | | EN-JA | | avg |
|---|---|---|---|---|---|---|---|---|---|
| | → | ← | → | ← | → | ← | → | ← | |
| Lample et al. (2018) (MUSE) | 0 | 0.15 | 0.11 | 0 | 0 | 0 | 34.52 | 3.56 | 4.79 |
| Artetxe et al. (2018a) (VecMap) | 0 | 0 | 0.11 | 0 | 0.08 | 0.28 | 43.80 | 31.64 | 9.49 |
| E and Zhou (2022) | 0 | 0 | 0 | 0 | 0 | 0.08 | 0 | 0 | 0.01 |
| Li et al. (2020) | 46.82 | 53.41 | 13.01 | 3.54 | 0.15 | 26.87 | 42.39 | 29.90 | 27.01 |
| Peng et al. (2021) | 0 | 0.30 | 0.11 | 0 | 23.98 | 31.91 | 43.50 | 31.95 | 16.47 |
| Proposed method (ECB best dim) | 48.36 | 23.41 | 15.57 | 3.94 | 26.36 | 33.12 | **50.04** | **39.14** | 29.99 |
| Proposed method (SCB dim 50) | 46.46 | 53.86 | 0.22 | 3.54 | **27.05** | 32.20 | 44.02 | 33.31 | 30.08 |
| Proposed method (SCB dim 100) | 45.51 | **55.96** | **16.35** | 3.28 | 26.82 | **33.40** | 48.85 | 38.08 | **33.53** |
| Proposed method (SCB dim 200) | 45.88 | 49.06 | 15.57 | **6.72** | 26.59 | 32.55 | 43.13 | 32.25 | 31.47 |
| Proposed method (SCB dim 300) | **48.43** | 53.26 | 15.57 | 3.67 | 26.82 | 32.62 | 46.10 | 35.12 | 32.70 |

Table 3: Results of Unsupervised method on the low-resource dataset of Section 4.2. We do 10 runs for each method and report the average score of the accuracies (%). For our proposed method, we use different dimensions of cross-lingual feature to do experiment. We show the best score of Embedding combination (ECB) method, and the score in four different dimensions of Similarity combination (SCB) method.

### 4.2 Datasets

We evaluate our method against baseline on the latest Wikipedia corpora. The reason why we do not use the famous dataset MUSE (Lample et al., 2018) is that we cannot get the corpus they used to train embeddings. So, we use FASTTEXT (Bojanowski et al., 2017) to train our own embeddings. However, the latest Wikipedia corpora is not on the same scale as the MUSE data. In order to ensure the comparability of models and dataset, and simulate low-resource situations, we reduce the size of corpus to match the results that the baselines on the MUSE dataset (For en:100%, vi:100%, th:10%, zh:10%, ja:30%). Our embeddings are trained based on this corpus. For the test dataset, we use the 1500 parallel lexicon of MUSE data.

### 4.3 Hyperparameter setting

We train our embedding using FASTTEXT with 5 epochs and 300 dims. For the hyperparameter $D$ in Problem (3), we set $D$ as 50. For the dimension $k$ in Problem (4), we experiment from 50 to 300 in increments of 50. For each language pair, we chose a different $\lambda$ between 0 and 1 to get the best results in Problem (7). When we do the initialization work, we only initialize 4000 words with the highest frequency in the monolingual corpus. For the iterative process, we only align the first 200000 words. We perform 10 runs for each language pair, and report the average accuracies. All the experiments are performed on a single Nvidia Titan X.

## 5 Experiment

In this section, we report the results obtained with our method. We first evaluate the dataset we trained in Section 5.1. Second, we present our main results in Section 5.2, thirdly we test the performance our cross-lingual feature in Section 5.3, then we do ablation tests in Section 5.4 to measure the contribution of each component and finally we compare the different initialization methods (3.2) in Section 5.5.

### 5.1 Comparison with MUSE dataset and our dataset

We compare the performance of our own trained dataset with the MUSE dataset (Lample et al., 2018). We use VecMap (Artetxe et al., 2018a) as an evaluation model. For each language pair, we experiment both two datasets on VecMap and the results are in Table 2. Our dataset restores MUSE dataset as far as possible in all experimental language pairs except for one direction of EN-JA (for EN-JA pair, we have a better result than MUSE). Our dataset offers the low-resource scenario and guarantees that our model is comparable to previous models.

### 5.2 Main results

We report the results in the dataset of we introduced in Section 4.2 in Table 3. As it can be seen, although the baselines succeed in some high-resource languages (EN-JA), they get the degradation in the challenging low-resource language pairs. In this case, our proposed method obtains the best results in all the language pairs. For the three language pairs on which the baselines are completely degraded (only 0.28% accuracy in the best pair), our method has made significant improvements in five experiments (16.13% at least and 54.23% at most). Peng et al. (2021) perform well on some language pairs (EN-ZH, EN-JA), but still fail on the low-resource pairs. For the state-of-the-art method (Li et al., 2020), we achieve the best score on every experiment. The average score of our method is 6.52% more than SOTA.

These results confirm the robustness of the proposed method. Our method converges to a good solution in all the low-resource and distant language pairs we experiment with. In addition to being more robust, our method also obtains better accuracies compared with the previous methods by a significant improvement in all challenging language pairs. Moreover, our method is not sensitive to hyperparameters. We can get similar results in different dimensions of cross-lingual features, and most of them perform a better result than all the baselines.

Meanwhile, our method is more efficient than VecMap (Artetxe et al., 2018a). For the iterative process of our method is based on VecMap, we compare between our method with VecMap. When we optimizing the initialization dictionary's quality, our method has a faster convergence rate (824 vs. 1346 iterations for EN-VI pair). It shows that a good initialization not only leads to a better accuracy, but also speeds up convergence.

### 5.3 Evaluation on cross-lingual feature

We test the performance of our cross-lingual feature in this section. We first compare cross-lingual features with pre-trained features and second-order of pre-trained features. For each language pair, we choose MUSE parallel dictionary (Lample et al., 2018), which contains 5000 aligned words as our dataset. We calculate the similarities between each aligned word using three kinds of features mentioned before. The results are shown in Table 4. As can be seen, the results of our cross-lingual features are significantly better than the pre-trained feature. For the second-order similarity, which represents a cross-lingual representation, our method is 0.024 more than it. Besides, our method is more efficient and not limited by computational complexity. The results show that our feature is cross-lingual.

For the case study, we choose 8 parallel pairs of high frequency words on EN-ZH language pair and calculate the similarity of cross-lingual features between each word pair. The results are shown in Figure 2. These similarities conform to our expectations. The words having the same meaning in different languages have similar representations which are shown in Figure 2 that elements on the main diagonal have a higher score of similarity of most word pairs. In particular, our feature can accurately distinguish four word pairs in all our eight tests. On the other hand, our feature exhibit good symmetry. For the similar meaning words <*one, first*>, their similarity with all the other words is similar which means the second and third rows in Figure 2 have similar results (the same for the second and third columns). Besides, for the word pairs that have a high similarity, their corresponding rows and columns have symmetry.

### 5.4 Ablation test

In order to better understand the role of each part in our proposed method, we do the ablation test to separately analyze the effect of cross-lingual features, pre-trained embeddings and normalization on initialization. We use the same setting with the best score in Table 6 (SCB dim 100) for the ablation test. The obtained results are shown in Table 5.

For our cross-lingual feature extraction method, we observe that the characterization ability of our feature is better than pre-trained embeddings, and the average of it exceeds the pre-trained embeddings by 7.47%. Moreover, our feature can be used

| Model | EN-VI | | EN-TH | | EN-ZH | | EN-JA | | avg |
|---|---|---|---|---|---|---|---|---|---|
| | → | ← | → | ← | → | ← | → | ← | |
| Pre-trained feature | 0.014 | 0.001 | 0.014 | 0.016 | 0.002 | 0.004 | 0.018 | 0.016 | 0.011 |
| Second-order similarity | 0.853 | 0.895 | 0.594 | 0.772 | 0.763 | 0.803 | 0.767 | 0.754 | 0.775 |
| Cross-lingual feature | 0.872 | 0.882 | 0.661 | 0.803 | 0.758 | 0.827 | 0.798 | 0.792 | 0.799 |

Table 4: Results of similarities on parallel dictionary. We calculate the similarities for each parallel words and report the average score of the similarities for each language pair. We experiment on pre-trained feature, second-order similarity of pre-trained feature and cross-lingual feature.

| Model | EN-VI | | EN-TH | | EN-ZH | | EN-JA | | avg |
|---|---|---|---|---|---|---|---|---|---|
| | → | ← | → | ← | → | ← | → | ← | |
| Full system | 45.51 | 55.96 | 16.35 | 3.28 | 26.82 | 33.40 | 48.85 | 38.08 | 33.53 |
| - Cross-lingual feature | 0.07 | 0.07 | 0.22 | 0.26 | 24.90 | 0.14 | 43.73 | 32.40 | 12.72 |
| - Pre-trained feature | 0 | 0 | 14.24 | 3.81 | 26.36 | 33.40 | 49.15 | 36.71 | 20.46 |
| - Normalization | 45.30 | 52.66 | 15.68 | 3.15 | 25.12 | 32.48 | 49.00 | 35.64 | 32.38 |

Table 5: Ablation test on the setting: SCB method, 100 dim of cross-lingual feature. We do 10 runs for each method and report the average score of the accuracies ($\%$) and the average accuracy score of all language pairs.

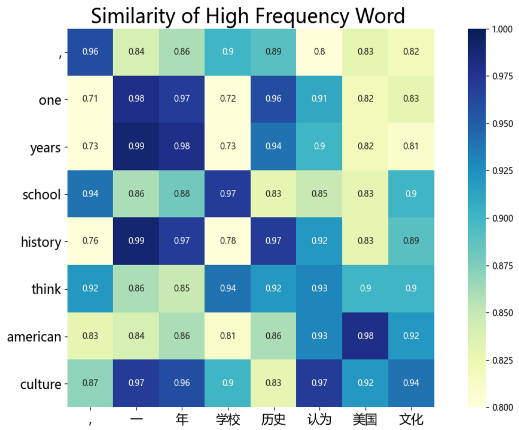

Figure 2: High frequency pairwise word similarity based on cross-lingual feature.

directly to initialize the dictionary. We do not need to calculate the second-order similarity, which reduces the problem size and computational complexity. Besides, the results show the complementary of these two kinds of features. They describe words from different dimensions. When we combined them using the two combination methods we proposed (ECB and SCB), they produce better results than either feature alone.

As for the normalization, the SCB method uses two kinds of features to compute similarity separately, and the results of regularization is not obvious. However, the ECB method uses concatenation, which is more sensitive.

### 5.5 Comparison with ECB and SCB method

In this section, we compare different initialization proposed in Section 3.2. As it can be seen in Table 6, the Similarity combination (SCB) method have a better result in most dimensions than Embedding combination (ECB) method. Besides, for each language pair, the best score among each dimension of SCB is better than ECB (expect EN-JA in Table 3). Especially, compared with ECB method, SCB method improves 31.95$\%$ in the reverse direction of EN-VI.

At the same time, the SCB method is less sensitive to the dimension parameters of cross-lingual features than ECB method. As we can see in Table 6, the SCB method produces stable and good results across different dimension settings, which shows that SCB method is more robust and can be adapted on more challenging problems.

| Model | EN-VI | | EN-TH | |
|---|---|---|---|---|
| | → | ← | → | ← |
| ECB dim 50 | 0 | 0.3 | 0 | 3.41 |
| ECB dim 100 | 0 | 1.65 | 0.11 | 3.94 |
| ECB dim 150 | 0.15 | 0.22 | 0.22 | 3.15 |
| ECB dim 200 | 0.28 | 0.75 | 14.35 | 3.28 |
| ECB dim 250 | 48.36 | 4.12 | 14.24 | 0.26 |
| ECB dim 300 | 0.73 | 23.41 | 15.57 | 0 |
| SCB dim 50 | 46.46 | 53.86 | 0.22 | 3.54 |
| SCB dim 100 | 45.51 | 54.38 | 16.24 | 3.28 |
| SCB dim 150 | 45.88 | 55.36 | 15.19 | 2.89 |
| SCB dim 200 | 45.88 | 49.06 | 15.57 | 6.72 |
| SCB dim 250 | 46.97 | 52.96 | 15.35 | 3.15 |
| SCB dim 300 | 48.43 | 53.26 | 15.57 | 3.67 |

Table 6: Results of different initialization on VecMap model. We do 10 runs for each method and report the average score of the accuracies (%). For each language pair, we initial VecMap with Embedding combination (ECB) method and Similarity combination (SCB) method in 6 different cross-lingual feature dimensions.

## 6 Related Work

Unsupervised bilingual lexicon induction (UBLI) is an important task of machine translation. The existing methods for unsupervised bilingual lexicon induction are divided into two directions. The first is based on statistic method, and the other is based on the pre-trained embeddings. Most of the methods follow the same procedure that is to find an initial solutions and then learning a mapping method between two embedding spaces. The key of these methods is finding an initial solution.

For statistical methods, Haghighi et al. (2008) induced translations for words by using a generative model based on canonical correlation analysis, which explains the monolingual lexicons in terms of latent matchings. Vulić et al. (2011) proposed a bilingual Latent Dirichlet Allocation model for finding translations of terms in comparable corpora without using any linguistic resources. E and Zhou (2022) proposed a method in a more mathematical way. Their Markov semantic model characterized the meaning of words with language-independent numerical fingerprints.

In recent years, most methods initialized seed dictionary based on pre-trained embeddings. These methods can be divided into three categories. The first category is using adversarial methods (Lample et al., 2018; Alvarez-Melis and Jaakkola, 2018; Xu et al., 2018). They trained a generator to find a mapping between two embedding spaces and a discriminator to distinguish the mapped source embedding from the target embedding. The second category is based on the structure of embedding space. Artetxe et al. (2018a) showed the fact that two equivalent words in different languages should have a similar distribution, and used the second-order similarity of pre-trained embeddings as an initialization of UBLI. The third category is based on a non-linear mapping method. Glavaš and Vulić (2020) removed the orthogonal constraint of the mapping method. Glavaš and Vulić (2020) proposed a non-linear mapping in the latent space of two independently pre-trained autoencoders.

All these methods only use one kind of feature. Different from their methods, we leverage both monolingual corpus and word-level pre-trained embeddings to get richer information and achieve better accuracy.

## 7 Conclusion

In this paper, we propose a method to extract cross-lingual features through monolingual corpora, combined with pre-trained embeddings in two kinds to initial UBLI. The experiments show that our method outperforms existing state-of-the-art methods on low-resource language pairs (EN-VI, EN-TH, EN-ZH, EN-JA). The ablation study demonstrates that the induced cross-lingual features have a complementary effect to pre-trained embeddings. Besides, we also offer a MUSE-equivalent dataset with monolingual corpora.

In the future, we will develop a more robust way of extracting cross-lingual features for lexicon induction. Extending UBLI to the phrase level is also a topic of interest.

## Acknowledgements

This work is supported by National Key RD Program of China under Grant 2020AAA0108000, National Natural Science Foundation of China under Grant U1908216 and Key R&D Program of Yunan under Grant 202203AA080004. The authors would like to thank the anonymous reviewers for their helpful comments.

## References

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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*, pages 2289–2294.

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

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Weinan E and Yajun Zhou. 2022. A mathematical model for universal semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1124–1132.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and e. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 57–63.

Eric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 526–533.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. 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 (Volume 1: Long Papers)*, pages 1234–1244.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: Hlt*, pages 771–779.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, pages 9–16.

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

Yanyang Li, Yingfeng Luo, Ye Lin, Quan Du, Huizhen Wang, Shujian Huang, Tong Xiao, and Jingbo Zhu. 2020. A simple and effective approach to robust unsupervised bilingual dictionary induction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5990–6001.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation.

Muhammad Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. 2020. Lnmap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2712–2723.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193.

Xutan Peng, Chenghua Lin, and Mark Stevenson. 2021. Cross-lingual word embedding refinement by $\ell_1$ norm optimisation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2690–2701.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 519–526.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.

Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Marco Saerens, and Kenji Fukumizu. 2013. Centering similarity measures to reduce hubs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 613–623.

Ivan Vulić, Wim De Smet, and Marie Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 479–484.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4407–4418.

Ivan Vulić and Marie Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1613–1624.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G Carbonell. 2019. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

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

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