A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction

Shuo Ren[†‡][*]**, Shujie Liu**[§]**, Ming Zhou**[§]**, Shuai Ma**[†‡]
[†]SKLSDE Lab, Beihang University, Beijing, China
[‡]Beijing Advanced Innovation Center for Big Data and Brain Computing, China
[§]Microsoft Research Asia, Beijing, China
[†]{shuoren,mashuai}@buaa.edu.cn [§]{shujliu,mingzhou}@microsoft.com

Abstract

Unsupervised bilingual lexicon induction is the task of inducing word translations from monolingual corpora of two languages. Recent methods are mostly based on unsupervised cross-lingual word embeddings, the key to which is to find initial solutions of word translations, followed by the learning and refinement of mappings between the embedding spaces of two languages. However, previous methods find initial solutions just based on word-level information, which may be (1) limited and inaccurate, and (2) prone to contain some noise introduced by the insufficiently pre-trained embeddings of some words. To deal with those issues, in this paper, we propose a novel graph-based paradigm to induce bilingual lexicons in a coarse-to-fine way. We first build a graph for each language with its vertices representing different words. Then we extract word cliques from the graphs and map the cliques of two languages. Based on that, we induce the initial word translation solution with the central words of the aligned cliques. This coarse-to-fine approach not only leverages clique-level information, which is richer and more accurate, but also effectively reduces the bad effect of the noise in the pre-trained embeddings. Finally, we take the initial solution as the seed to learn cross-lingual embeddings, from which we induce bilingual lexicons. Experiments show that our approach improves the performance of bilingual lexicon induction compared with previous methods.

1 Introduction

Bilingual lexicon induction (BLI) is an important task of machine translation and becomes an essential part of recent unsupervised machine translation approaches (Lample et al., 2018; Artetxe et al., 2018c; Marie and Fujita, 2018; Ren et al., 2019; Artetxe et al., 2019). Previous methods for BLI are mostly based on unsupervised cross-lingual word embeddings (Zhang et al., 2017; Artetxe et al., 2017; Conneau et al., 2017; Artetxe et al., 2018b; Xu et al., 2018; Hoshen and Wolf, 2018; Alvarez-Melis and Jaakkola, 2018), the goal of which is to find a mapping function, typically a linear transformation (Mikolov et al., 2013), to map the source embeddings into the target embedding spaces. To do this, they first build a seed dictionary (known as the initial solution) with different methods and then learn the optimal mapping function that fits the seed dictionary. Based on the mapping function, a new dictionary of higher quality is inferred from the cross-lingual word embeddings by finding nearest neighbors in the target embedding space. With the new dictionary, the mapping function is further refined to fit it. The inference of the dictionary and the refinement of the mapping function are iteratively done until the final convergence. During the whole procedure, the initialization stage is important and heavily focused in previous work.

Previous methods for finding the initial solution fall into three categories. The first one is heuristic rules such as treating identical words as the seed (Artetxe et al., 2017), but this kind of method is restricted to languages sharing the alphabet. The second category is adversarial methods (Zhang et al., 2017; Conneau et al., 2017; Xu et al., 2018; Alvarez-Melis and Jaakkola, 2018), but suffering from the drawbacks of generative adversarial models, i.e., the sensitivity of hyper-parameters, long training time, etc. The third category is structure-based methods (Artetxe et al., 2018b; Hoshen and Wolf, 2018), which is more flexible and robust than other categories, and achieve the state-of-the-art BLI performance. In Artetxe et al. (2018b), they first compute a similarity matrix of all words in the vocabulary, and then represent each word with the distribution of the similarity values, while in Hoshen and Wolf (2018), they project the word

[*] Contribution during internship at MSRA.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3476–3485
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

vectors to the top 50 principal components of the embedding spaces. After that, both of them directly use the word representation of two languages to retrieve the initial bilingual lexicons by computing the cosine distances of source and target word representations. However, directly finding word alignments from scratch has some demerits. (1) The information that a word can provide is limited and independent of each other. (2) According to our observation, there is some noise in the pre-trained embeddings even for high-frequency words so that the initial word alignments derived from them are not accurate. Those mistakes in the initial word-level alignments can hurt the performance in the following iteration steps.

To solve those issues, we propose a novel graph-based coarse-to-fine paradigm to generate initial solutions for learning cross-lingual word embeddings, from which we induce bilingual lexicons. Specifically, given source and target languages, our method first uses pre-trained monolingual embeddings to construct a graph for each language, with the vertices representing different words, so that the mutual relationship between words is preserved. Next, we use the Bron–Kerbosch algorithm (Akkoyunlu, 1973) to extract cliques (a subset of vertices in which every two distinct vertices are adjacent) in the source and target graphs. After that, we calculate the clique embeddings and map the cliques from two graphs. We then treat the central words of the aligned cliques as the seeds to learn the mapping of the two word embedding spaces.

Our contributions are threefold. (1) By building word graphs, we leverage the clique-level information extracted from them. The cliques cluster similar words and assemble their mutual relationship of them, providing richer and more accurate information. (2) We propose the coarse(clique extraction)-to-fine(seed induction) procedure for the BLI task, which effectively reduces the bad effect of the noise in the pre-trained embeddings; (3) We improve the BLI performance on the MUSE dataset with our method, even compared with strong baselines.

2 Background

Unsupervised bilingual lexicon induction (BLI) is the task of inducing word translations from monolingual corpora of two languages. Recently proposed methods follow the same procedure, i.e., first learning cross-lingual embeddings in an unsupervised way (§2.1) and then inducing bilingual lexicons from the embedding spaces (§2.2).

2.1 Unsupervised Cross-lingual Embeddings

Previous methods for learning cross-lingual embeddings can be roughly divided into two categories (Ormazabal et al., 2019), i.e., mapping methods and joint learning methods. As the second category, the skip-gram (Luong et al., 2015) for example, requires bilingual corpus during training, current methods for unsupervised cross-lingual embeddings mainly fall into the first category. Given pre-trained monolingual embeddings of two languages, the mapping methods try to map the source and target embedding spaces through a linear transformation (Mikolov et al., 2013) $\mathbf{W} \in \mathbf{M}_{d\times d}(\mathbb{R})$, where $\mathbf{M}_{d\times d}(\mathbb{R})$ is the space of $d \times d$ matrices of real numbers and d is the dimension of the embeddings. Based on that, Xing et al. (2015) propose to constrain $\mathbf{W}$ to be orthogonal, i.e., $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$, and Conneau et al. (2017) find this is a Procrustes problem which advantageously offers a closed-form solution obtained from singular value decomposition (SVD) of $\mathbf{YX}^\top$ as follows:

$$\begin{aligned} \mathbf{W}^* &= \arg\min_{\mathbf{W}} ||\mathbf{WX} - \mathbf{Y}||_F = \mathbf{UV}^\top, \\ &\text{with } \mathbf{U\Sigma V}^\top = \text{SVD}\,(\mathbf{YX}^\top) \end{aligned} \tag{1}$$

where $\mathbf{X}$ and $\mathbf{Y} \in \mathbf{M}_{d\times n}(\mathbb{R})$ consist of the embeddings of the bilingual lexicons $\{x_i, y_i\}_{i=1}^n$ in the seed dictionary.

Therefore, there are two steps to learn unsupervised cross-lingual embeddings. The first step is to find an initial solution (also known as the seed dictionary), and the second one is to obtain the desired $\mathbf{W}$ according to Eq. (1). The above two steps can be iteratively done, by inducing new seed dictionary from the learned cross-lingual embeddings with the method introduced next, and using the new dictionary to refine the matrix $\mathbf{W}$ (known as the "**refinement**" process in some literature).

The first step, i.e., finding the initial solution, is crucial because it decides the direction of the following iteration. Loads of previous work are devoted to finding good initial solutions with different methods, as is described in §1. But their methods only exploit word-level information, which is limited and may be inaccurate due to the noise in pre-trained monolingual embeddings, leading to mistakes in the initial word-level alignments. Therefore, we propose a novel graph-based coarse-to-fine paradigm to find the initial solution of higher quality, leveraging clique-level information which we think is richer and more accurate.

2.2 Bilingual Lexicon Induction

Based on the learned cross-lingual embeddings, bilingual lexicons can be induced from the mapped spaces via the nearest neighbor (NN) method by calculating the cosine distance of the mapped source embeddings and the target embeddings. However, this method suffers from the "hubness" problem (Dinu et al., 2014) such that some target words appear as the nearest neighbors of many source words. To mitigate this problem, alternatives of the distance function have been proposed, such as invsoftmax (Smith et al., 2017), CSLS (Conneau et al., 2017) and margin-based scores (Artetxe and Schwenk, 2018). Among them, CSLS, as a special case of margin-based scores, is widely used in the SOTA embedding-based BLI methods. Formally, CSLS calculates the distance between the mapped and the target embeddings as follows:

$$\mathrm{CSLS}(\mathbf{Wx}, \mathbf{y}) = 2\cos(\mathbf{Wx}, \mathbf{y}) - r_{\mathrm{T}}(\mathbf{Wx}) - r_{\mathrm{S}}(\mathbf{y}) \tag{2}$$

where

$$r_{\mathrm{T}}(\mathbf{Wx}) = \frac{1}{K} \sum_{y \in \mathcal{N}_{\mathrm{T}}(\mathbf{Wx})} \cos(\mathbf{Wx}, \mathbf{y}) \tag{3}$$

is the mean similarity of a source embedding $\mathbf{x}$ to its K target neighborhoods ($\mathcal{N}_{\mathrm{T}}(\mathbf{Wx})$). Similarly, $r_{\mathrm{S}}(\mathbf{y})$ is the mean similarity of a target embedding $\mathbf{y}$ to its neighborhoods.

3 Methodology

As is mentioned before, recent work on bilingual lexicon induction (BLI) is mostly based on unsupervised cross-lingual embeddings, whose key point is to find initial solutions to learn the mapping function. However, previous methods find initial solutions just based on word-level information, which may be limited and inaccurate due to the noise in pre-trained monolingual embeddings. Therefore, we exploit the information provided by word cliques and figure out a coarse-to-fine procedure to denoise and find the initial solution of higher quality. Based on that, we learn the cross-lingual embeddings and induce word translations.

As shown in Figure 1, our method for BLI can be roughly divided into several steps. Given the source and target languages, we first build a graph for each language. The graph vertex represents the word. Next, we extract word cliques from the graphs and map the cliques of two languages in an unsupervised way. Then, we induce the seed dictionary from the bilingual cliques by choosing the respective central words of the aligned cliques. After that, we learn cross-lingual embeddings with the help of the induced seed dictionary. The above steps can be iteratively done until the final convergence. By building word graphs, we can use the clique-level information which is richer and more accurate than what a single word provides. Besides, the whole coarse-to-fine procedure also reduces the bad effect of the noise in the pre-trained embeddings, because the clique-level alignment (coarse) is more accurate at the beginning and the word alignments inferred from it (fine) are more reasonable. We will next introduce each step.

3.1 Word Graph Construction

Given the pre-trained monolingual embeddings, we can derive an edge-weighted graph from them by regarding words as the vertices and their similarities as edges. Formally, the graph is

$$G =< V, E > \tag{4}$$

where V is the vertex set (vocabulary of each language) and E is the edge set. The edges are built with monolingual embedding similarities. For example, for language x, to define the edges, we first get the word-to-word similarity matrix $\mathbf{M}$ with

$$\mathbf{M}_{i,j} = \begin{cases} \mathrm{CSLS}(\mathbf{x}_i, \mathbf{x}_j), & i \neq j \\ 0, & i = j \end{cases} \tag{5}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the normalized embeddings of two words respectively. We set the main diagonal elements to zero to avoid self-loop. Theoretically, there is one edge between any two arbitrary words with the edge weight to be $\mathbf{M}_{i,j}$, but if the weight of an edge is too small, it will provide little information and introduce a lot of noise. Therefore, we prune these non-informative edges with $\mathbf{M}_{i,j}$ less than a threshold of θ. Meanwhile, the pruning greatly reduces the computation time of the next step. We build two graphs G_x and G_y for two languages x and y in this way respectively.

3.2 Clique Extraction and Mapping

Different from previous methods, we infer the initial solution not using word-level information but from word cliques, which we think is richer and more accurate. Following Wang et al. (2016), the

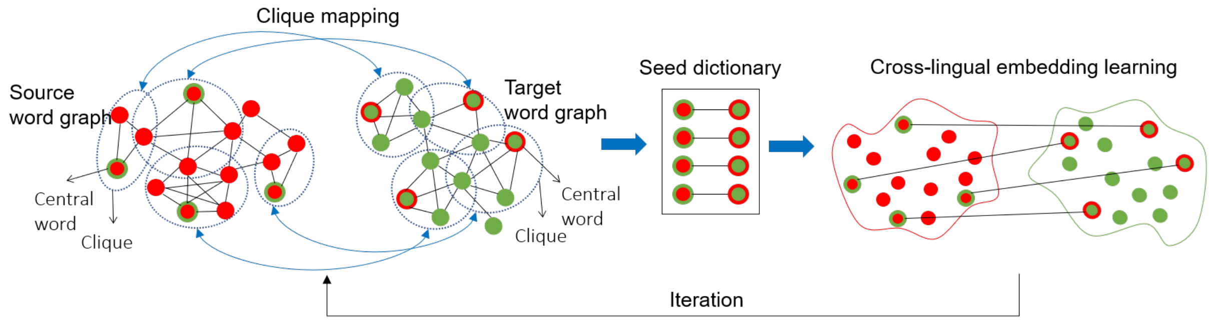

Figure 1: Overview of our method. In each iteration, based on the word graphs, we first map the cliques of two languages in an unsupervised way, and then infer the seed dictionary to learn cross-lingual word embeddings.

"clique" here means a maximum complete subgraph where every two distinct vertices in the clique are adjacent. Extracting cliques from a given graph is a nontrivial problem and is shown to be NP-complete (Karp, 1972). In this paper, we adopt Bron-Kerbosch (BK) algorithm (Akkoyunlu, 1973) with pivoting (Johnston, 1976) to extract the cliques from a given graph. Having extracted the word cliques of two languages, we calculate clique embeddings by averaging the embedding vectors of all words in each clique. We choose the word whose embedding is closest to its clique embedding as the central word of each clique. After that, we follow Artetxe et al. (2018b) to map the cliques of two languages in a fully unsupervised way, i.e. to learn cross-lingual clique embeddings.

We use the clique extraction rather than clustering methods because (1) a word may fall into different categories because of polysemy, which can be well modeled by the cliques, and (2) the BK algorithm is much more efficient than clustering.

3.3 Seed Dictionary Induction

§3.2 maps the clique embeddings of two languages into the same space so that we can retrieve aligned cliques. For each source clique, we choose the nearest target clique according to the CSLS similarity score calculated by Eq. (2). Remember that we have chosen the central word for each clique after the clique extraction in §3.2, so the seed dictionary inferring process is simply picking the central words of the aligned cliques just as shown in Figure 1. Note that we remove the duplication of seed word pairs in this process.

3.4 Cross-lingual Embedding Learning

Based on the initial solution (known as the seed dictionary), we then learn cross-lingual word embeddings following the Procrustes and refinement process introduced in §2.1. After obtaining the learned cross-lingual word embeddings, we rebuild the word graphs with the help of them and iterate the whole process again until the final convergence as shown in Figure 1.

Previously methods used a single matrix $\mathbf{W}$ as transformation function between the embedding spaces of two languages, based on the assumption that the embedding spaces of different languages are isomorphic (Mikolov et al., 2013). However, this is doubtful because the isomorphic assumption may not hold all the time (Søgaard et al., 2018). Fortunately, the cliques we extracted naturally provide good local features for us, because they are usually much different from each other in meanings, which enables us to investigate alternatives to a single mapping matrix $\mathbf{W}$. Therefore, after the final iteration, we divide all the cliques into K groups via clustering, i.e., $\{L_i\}_{i=1}^{K}$, and train an individual matrix $\mathbf{W}_i$ for each of them. We denote this process as **"group mapping"**. Each $\mathbf{W}_i$ is initialized with the learned $\mathbf{W}$ and fine-tuned as

$$\mathbf{W}_i = \underset{\mathbf{W}_i}{\arg\min} ||\mathbf{W}_i\mathbf{X}_i - \mathbf{Y}_i||_{\mathrm{F}}, \text{ s.t. } \mathbf{W}_i^{\top}\mathbf{W}_i = \mathbf{I} \tag{6}$$

where $\mathbf{X}_i$ and $\mathbf{Y}_i$ are the embedding matrices of words belonging to L_i. We divide each word into the group closest to its word embedding. The whole training procedure is shown in Algorithm 1.

3.5 Inference

After the training, we can obtain the renewed word graphs of both languages as well as their cliques, and get a set of group mapping matrices $\{\mathbf{W}_i\}_{i=1}^{k}$. During the inference, for each source word x, we first find its closest clique C_s by calculating the similarities of x's embeddings to all clique embeddings. Next, we retrieve the group L_s that C_s belongs to, and choose the corresponding $\mathbf{W}_s$. Then,

Algorithm 1: Training procedure of the proposed graph-based coarse-to-fine method.

```
Input: Monolingual embeddings of two languages $\mathbf{X}$, $\mathbf{Y}$
Output: Multiple local mapping matrices $\{\mathbf{W}_i\}_{i=1}^{m}$
while not convergence do
    Build the word graphs $G_x$ and $G_y$ by calculating
      the embedding similarities of each language.
    Extract cliques $\{C_i^x\}_{i=0}^{m}$ and $\{C_j^y\}_{j=0}^{n}$ from each
      graph using the Bron-Kerbosch algorithm.
    Calculate the clique embeddings by averaging the
      embeddings of all the words belonging to it.
    Map the source and target cliques with the method
      of Artetxe et al. (2018b).
    Build seed dictionary with the central words of the
      aligned cliques.
    Do the Procrustes and refinement iteration described
      in §2.1 and learn the mapping matrix $\mathbf{W}$.
    Renew the embeddings of the source language as
      $\mathbf{X} := \mathbf{W}\mathbf{X}$.
Divide $\{C_i^x\}_{i=0}^{m}$ into $K$ groups via clustering. Initialize
  $\{\mathbf{W}_i\}_{i=0}^{K}$ with $\mathbf{W}$.
Fine-tune each $\mathbf{W}_i$ according to Eq. (6) and do the
  refinement.
return $\{\mathbf{W}_i\}_{i=0}^{K}$
```

we retrieve the translation of x by calculating the CSLS score of $\mathbf{W}_s\mathbf{x}$ and each target embedding $\mathbf{y}$, similar to Eq. (2) introduced in §2.2.

4 Experiment

4.1 Dataset

Bilingual lexicon induction (BLI) measures the word translation accuracy in comparison to a gold standard. We report results on the widely used MUSE dataset (Conneau et al., 2017). This dataset consists of monolingual fastText (Bojanowski et al., 2017) embeddings of many languages and dictionaries for many language pairs divided into training and test sets. The evaluation follows the setups of Conneau et al. (2017).

4.2 Implementation Details

4.2.1 Pre-processing

We choose the top 10,000 word embeddings to build word graph because the monolingual embeddings of low-frequency words may be trained insufficiently. The embeddings are normalized following Artetxe et al. (2018b). Specifically, we first apply length normalization to the embeddings, and then mean center each dimension. After that, we do length normalization again to ensure the word embeddings have a unit length.

4.2.2 Clique Extraction

An efficient algorithm for clique extraction is the Bron-Kerbosch (BK) algorithm, which is a recursive backtracking algorithm that searches for all maximal cliques in a given graph G. The pruning operation described in §3.1 makes the word graph a sparse graph, for which the BK algorithm can be made to run in time $O(dn3^{d/3})$ (Eppstein and Strash, 2011), where n is the number of vertexes in G, and d is the degeneracy [1] of the graph. We choose a public efficient C implementation of BK algorithm [2], and only extract the cliques that contain no less than three words. According to our observation, the cliques can be extracted within several seconds with this code.

4.2.3 Clique and Word Embedding Mapping

In our experiment, the clique embeddings of two languages are mapped with the method proposed by Artetxe et al. (2018b). We use their public code to finish this step. We initialized W with a random orthogonal matrix. After building the seed dictionary, we first solve the Procrustes problem (Eq. (1)), followed by the refinement process.

4.3 Main Results

4.3.1 Baselines

We choose several supervised and unsupervised methods to be our baselines. The supervised baselines include: (1) The iterative Procrustes method proposed by Smith et al. (2017); (2) The multi-step framework proposed by Artetxe et al. (2018a); (3) a geometric method proposed by Jawanpuria et al. (2019). The unsupervised baselines include (1) MUSE proposed by Conneau et al. (2017), which is a GAN based method followed by a refinement process; (2) a Wasserstein GAN based method combined with distribution matching and back translation, proposed by Xu et al. (2018); (3) a method proposed by Alvarez-Melis and Jaakkola (2018) that views the mapping problem as optimal transportation and optimize the Gromov-Wasserstein distance between embedding spaces; (4) A robust self-learning method proposed by Artetxe et al. (2018b), which leverages the intra-linguistic word similarity information to infer initial solutions, followed by a self-learning iteration; (5) A non-adversarial method proposed by Hoshen and Wolf

[1] In graph theory, a k-degenerate graph is an undirected graph in which every subgraph has a vertex of degree $\leq k$

[2] https://github.com/aaronmcdaid/MaximalCliques

Method	en-fr		en-de		en-es		en-it		en-ru		en-zh	
	→	←	→	←	→	←	→	←	→	←	→	←
						Supervised						
(Smith et al., 2017)	81.1	82.4	73.5	72.4	81.4	82.9	43.1	38.0	51.7	63.7	42.7	36.7
(Artetxe et al., 2018a)	80.5	83.1	73.5	73.5	80.5	83.8	**61.3**	**39.6**	50.5	67.3	32.3	43.4
(Joulin et al., 2018)	**83.3**	84.1	**79.1**	76.3	**84.1**	**86.3**	-	-	**57.9**	67.2	45.9	**46.4**
(Jawanpuria et al., 2019)	82.1	**84.2**	74.9	**76.7**	81.9	85.5	-	-	52.8	**67.6**	**49.1**	45.3
						Unsupervised						
(Conneau et al., 2017)	82.3	81.1	74.0	72.2	81.7	83.3	77.4	76.1	44.0	59.1	32.5	31.4
(Xu et al., 2018)	77.9	75.5	69.3	67.0	79.5	77.8	72.6	73.4	-	-	-	-
(Alvarez-Melis and Jaakkola, 2018)	81.3	78.9	71.9	72.8	81.7	80.4	78.9	75.2	45.1	43.7	-	-
(Artetxe et al., 2018b)	82.3	83.6	75.1	74.3	82.3	84.7	78.8	79.5	49.2	**65.6**	-	-
(Hoshen and Wolf, 2018)	82.3	**84.1**	74.7	73.0	82.1	84.1	77.9	77.5	47.5	61.8	-	-
Ours (without GM)	82.7	83.4	**75.5**	75.7	82.6	84.8	78.6	79.5	48.9	63.9	38.1	35.2
Ours (with GM)	**82.9**	83.9	75.3	**76.1**	**82.9**	**85.3**	**79.1**	**79.9**	**49.7**	64.7	**38.9**	**35.9**

Table 1: Precision@1 for the MUSE BLI task. All baselines leverage CSLS to be the retrieve metric during inference except for Xu et al. (2018) which uses cosine similarity. The bold numbers indicate the best results of supervised and unsupervised methods. "GM" means applying the group mapping technique described in §3.4.

(2018), which uses PCA-based alignment to initialize and iteratively refine the alignment.

4.3.2 Results of Common Languages

We report the result of the BLI task on the MUSE dataset (Conneau et al., 2017). The language pairs we choose are French (fr), German (de), Spanish (es), Italian (it), Russian (ru), Chinese (zh) from and to English(en), as shown in Table 1.

From Table 1, we find that our proposed method significantly outperforms previous methods on nearly all directions, especially on en-de and en-zh pairs, with the improvements of 2 to 6 points compared with previous state-of-the-art unsupervised approaches. The results on some language pairs such as en-fr, en-de and en-es are remarkably competitive with strong supervised methods.

We also see that for distant languages, i.e., en-ru and en-zh, our method achieves good results, on which some unsupervised baselines fail to converge. However, the results are still far lagging behind the supervised methods, indicating that the seed dictionaries built with our method may not be perfect for these distant languages. This may root in the original diversified training data of the monolingual embeddings on those pairs. Even so, we still significantly outperforms the MUSE (Conneau et al., 2017) for the en-ru and en-zh pairs.

4.3.3 Results of Morphologically Rich Languages

We also list results of some morphologically rich languages, i.e., Finnish (fi), Polish (pl) and Turkish (tr) in Table 2, which are selected by Søgaard et al. (2018). They find that these languages are different in morphological traits from commonly benchmarked languages which are morphological poor isolating or exclusively concatenating languages. For these languages, Søgaard et al. (2018) leverage identical tokens in both languages as the seeds (Artetxe et al., 2017), followed by the Procrustes solution plus the refinement process, which generates relatively good results. We compare our results with the supervised method, i.e., use 5k dictionary to start up followed by Procrustes + refinement, MUSE (Conneau et al., 2017) and Søgaard et al. (2018) on these languages.

Method	en-fi		en-pl		en-tr	
	→	←	→	←	→	←
			Supervised			
5k+Pro.+Ref.	47.3	59.5	58.2	66.9	46.3	59.2
			Unsupervised			
(Conneau et al., 2017)	0.1	59.8	53.9	0.0	45.4	0.0
(Søgaard et al., 2018)	45.0	59.1	57.3	66.7	45.4	**61.4**
Ours (without GM)	47.1	59.2	59.7	68.4	50.2	59.7
Ours (with GM)	**48.1**	**60.4**	**60.8**	**69.0**	**51.4**	60.9

Table 2: Precision@1 for the MUSE BLI task of morphologically rich languages. The bold numbers indicate the best results of all methods. Pro.: Procrustes; Ref.: Refinement.

From the table, we see that the GAN-based method (MUSE) fails to give good results of some directions, maybe due to its unstable training. Using identical tokens as the seed gives good results (Søgaard et al., 2018) and compares with the supervised method. Our method performs well on these morphologically rich languages, and even outperforms the supervised method. We also conduct experiments on other morphologically rich languages such as Estonian, Greek, and Hungarian, but fail to converge.

4.3.4 Effect of Group Mapping

From Table 1 and Table 2, we also find that leveraging the group mapping (GM, §3.4) contributes to bilingual lexicon induction, especially for some distant languages such as en-ru, en-zh, and morphologically rich languages, with the improvement from 0.7 to 1.2 points. This result indicates the assumption that the embedding spaces of different languages are isomorphic may only hold locally. With the help of the cliques we extracted, we can find those locality features via clustering.

4.4 Sensitivity to Hyper-parameters

Notice that our method depends on three major hyper-parameters: (1) the number of words N we use to build word graphs; (2) the threshold θ to prune the edges in the graphs; (3) the number of iterations I we do. In this subsection, we discuss the impact of these hyper-parameters on the BLI results, taking en2fr as an example. We depict the precision@1 on different hyper-parameter settings in Figure 2.

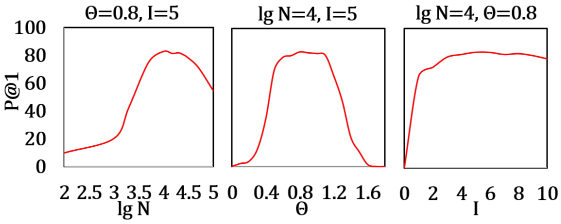

Figure 2: Influence of the hyper-parameters.

From the figure, we find that the performance of our method is sensitive to the choice of N and θ. If N is too small, the cliques extracted cannot reach agreement semantically across different languages because of the sparsity of semantic units. If N is too large, the improperly trained low-frequency word vectors will impair the performance too. As for θ, if the threshold is too small, then much noise will be introduced into the word graphs, not only reducing the quality of extracted cliques but increasing the execution time of the BK algorithm. For I, we find that the performance improves fast when I is increased from 0 to 2, but reaches convergence at 5. Too many iterations hurt the performance because, at this time, the seed dictionary inferred from the mapped cliques is redundant.

4.5 Influence to Unsupervised MT

It has been shown that BLI can benefit unsupervised machine translation (MT) (Lample et al., 2018; Marie and Fujita, 2018; Ren et al., 2019) by building Statistical Machine Translation (SMT) with the induced bilingual lexicons and language models as SMT features, followed by an iterative back-translation process. In this part, we will discuss the influence of different bilingual lexicon induction methods (Conneau et al., 2017; Artetxe et al., 2018b) to the performance of the initial SMT model, and report the BLEU scores[3] on *newstest2014* en-fr and en-de tasks in Table 3. Note that we do not do the subsequent iterative back-translation process. From the table, we see that the performance of unsupervised SMT is restricted to the quality of BLI results. As our method provides better word translations, the initial SMT models benefit from ours accordingly.

BLI Method	en2fr	fr2en	en2de	de2en
MUSE	11.74	15.34	8.14	11.03
VecMap	13.04	16.40	9.12	11.98
Ours	13.91	17.21	10.24	12.41

Table 3: BLEU of initial unsupervised SMT. The SMT features are word translation tables inferred from different BLI methods and pre-trained language models.

5 Case Study

5.1 Extracted Cliques

In this part, we give some examples of the English cliques extracted with our method, as listed in Table 5. From the table, we see that our method can extract reasonable cliques containing words that share similar meanings. Each clique can be regarded as a semantic unit, which is more explicit than the PCA-based initialization method (Hoshen and Wolf, 2018) where they represent the semantic units with a fixed number of principal components. An interesting phenomenon is that "May" is not in the fifth clique which groups all the words of months. This is because, in this dataset, all the words are lower-cased so that "may" is also a modal verb. Besides, we observe the extracted cliques of other languages and find they are also reasonable, which are not listed here due to space limitation.

[3] Tested by *multi-bleu.pl*.

en	fr			zh		
	MUSE	VecMap	Ours	MUSE	VecMap	Ours
and	part(share)	établir(establish)	et(and)	也(too)	/	和(and)
his	n	matin(morning)	lui(him)	此(now)	第六(sixth)	他(he)
south	un (a)	avait(had)	ouest(west)	台北(Taipei)	(prize)	北(north)
august	flotte(fleet)	mars(march)	mars (march)	电影(film)	第五(fifth)	三月(march)
build	paris(Paris)	seule(alone)	faire(make)	用作(used as)	了解(understand)	形成(form)

Table 4: Examples of seeds produced with different methods. Inside the brackets is the interpretation of the words.

id	words
1	, . **-**) (
2	**and** also both well addition additionally besides
3	**his** himself him he her
4	northeastern west **south** southeastern southeast east southwest northeast northwest southwestern north
5	january march **august** july september october june april december november february
6	science **scientists** scientific biology mathematics physics chemistry sciences
...	...

Table 5: Examples of English cliques extracted from the word graph in the first iteration. The bold words are the central words in their respective cliques.

5.2 Seed Dictionary

To demonstrate that our method can produce good initial solutions for learning cross-lingual embeddings, in this part, we give an example of the seed dictionary inferred during the first iteration with our method, compared with that inferred by MUSE (Conneau et al., 2017) and VecMap (Artetxe et al., 2018b). The language pairs we choose are en-fr and en-zh, as listed in Table 4. From the table, we find that our method produces initial solutions with higher quality. This is because our coarse-to-fine process can effectively filter out the noise from the start. Notice that the initial solution produced by MUSE in the first iteration is not good, which may be because the GAN based method is not stable enough at the beginning of the training.

6 Related Work

Bilingual lexicon induction (BLI) is an important task of machine translation. Recent methods for bilingual lexicon induction are mostly based on unsupervised cross-lingual word embeddings (Zhang et al., 2017; Artetxe et al., 2017; Conneau et al., 2017; Artetxe et al., 2018b; Xu et al., 2018; Hoshen and Wolf, 2018; Alvarez-Melis and Jaakkola, 2018). They follow the same procedure that is first building initial solutions (a seed dictionary) and then learning a mapping function between the two word embedding spaces. During inference, for a given source word, they find the target word via the nearest neighbors search by calculating the distance of the mapped source embedding and all target word embeddings. The main focus of the previous methods is how to find the initial solution, which is the most important part.

Their methods can be divided into three categories according to the way of finding the initial solution. The first category is using heuristic rules such as treating identical words as the seed (Artetxe et al., 2017), but this kind of method is restricted to languages sharing the vocabulary or at least the notation of numbers. The second category is adversarial methods (Zhang et al., 2017; Conneau et al., 2017; Xu et al., 2018; Alvarez-Melis and Jaakkola, 2018). They train a generator to finish mapping between the two word embedding spaces, and a discriminator to distinguish the mapped embeddings from the target embeddings. However, they suffer from the drawbacks of generative adversarial models, i.e., the sensitivity of hyper-parameters, long training time and lack of interpretability (Hoshen and Wolf, 2018). The third category is structure-based methods, which achieve the state-of-the-art performance on BLI. They either leverage the intra-linguistic word similarity information (Artetxe et al., 2018b) or principal components of monolingual word embeddings (Hoshen and Wolf, 2018), but their methods infer initial solutions just based on word-level information which is limited and prone to contain some noise due to the insufficient training of pre-trained embeddings. Different from their methods, ours leverages clique-level information which is richer and more accurate, and uses a coarse-to-fine procedure to reduce the adverse effect of the noise mentioned above.

7 Conclusion

In this paper, we propose a novel graph-based coarse-to-fine paradigm for unsupervised bilingual lexicon induction. Our method uses clique-level information and reduces the bad effect of noise in the pre-trained embeddings. The experiments show that our method can significantly improve the bilingual word induction performance after several iterations compared with strong baselines, even for distant language pairs. In the future, we will consider combining our method with Graph Neural Networks to update the word graphs we build.

Acknowledgments

This work is supported in part by National Key R&D Program of China AAA0102301, and NSFC 61925203 & U1636210 & 61421003.

References

Eralp Abdurrahim Akkoyunlu. 1973. The enumeration of maximal cliques of large graphs. *SIAM Journal on Computing*, 2(1):1–6.

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on EMNLP*, pages 1881–1890.

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 ACL (Volume 1: Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018c. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on EMNLP*, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of ACL*.

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

David Eppstein and Darren Strash. 2011. Listing all maximal cliques in large sparse real-world graphs. In *International Symposium on Experimental Algorithms*, pages 364–375. Springer.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on EMNLP*, pages 469–478.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: a geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120.

HC Johnston. 1976. Cliques of a graph-variations on the bron-kerbosch algorithm. *International Journal of Computer & Information Sciences*, 5(3):209–238.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on EMNLP*, pages 2979–2984.

Richard M Karp. 1972. Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Springer.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on EMNLP*, pages 5039–5049.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*.

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

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of ACL*.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. In *Thirty-Three AAAI Conference on Artificial Intelligence*.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

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

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, and Masao Utiyama. 2016. A bilingual graph-based semantic model for statistical machine translation. In *IJCAI*, pages 2950–2956.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of NAACL: Human Language Technologies*, pages 1006–1011.

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 EMNLP*, 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 ACL (Volume 1: Long Papers)*, pages 1959–1970.