DM-BLI: Dynamic Multiple Subspaces Alignment for Unsupervised Bilingual Lexicon Induction

Anonymous ACL submission

Abstract

Unsupervised bilingual lexicon induction (BLI) task aims to find word translations between languages and has achieved great success in similar language pairs. However, related works mostly rely on a single linear mapping for language alignment and fail on distant or low-resource language pairs, achieving less than half the performance observed in rich-resource language pairs. In this paper, we introduce **DM-BLI**, a **D**ynamic **M**ultiple subspaces alignment framework for unsupervised **BLI**. DM-BLI improves language alignment by utilizing multiple subspace alignments instead of a single mapping. We begin via unsupervised clustering to discover these subspaces in source embedding space. Then we identify and align corresponding subspaces in the target space using a rough global alignment. DM-BLI further employs intra-cluster and inter-cluster contrastive learning to refine precise alignment for each subspace pair. Experiments conducted on standard BLI datasets for 12 language pairs (6 rich-resource and 6 low-resource) demonstrate substantial gains achieved by our framework. We release our code to facilitate the community.

1 Introduction

Unsupervised bilingual lexicon induction (BLI) has shown to be a key multilingual NLP task to align cross-lingual word embeddings (CLWE) (Mikolov et al., 2013a; Ruder et al., 2019) and bridge lexical gap between low-resource languages (Eder et al., 2021; Marchisio et al., 2022).

Existing BLI approaches can be roughly divided into two categories: mapping-based methods (Conneau et al., 2017; Artetxe et al., 2018; Ren et al., 2020; Li et al., 2022) and generation-based methods (Gonen et al., 2020; Ghazvininejad et al., 2023; Li et al., 2023). Mapping-based methods aim to align monolingual embeddings from various languages into a shared CLWE space via linear or non-linear projections. Generation-based methods

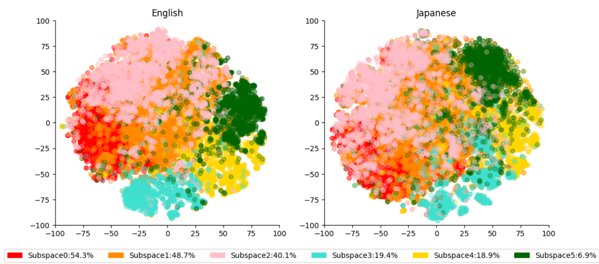

Figure 1. t-SNE visualization of the clustered monolingual word embedding in a distant language pair of English (left) and Japanese (right). Different colors represent different subspaces. With a global orthogonal mapping from English to Japanese, BLI accuracies for subspaces 0-5 are 54.3%, 48.7%, 40.1%, 19.4%, 18.9% and 6.9%, respectively.

leverage the machine translation capacities of large language models (LLMs) (Briakou et al., 2023) to directly generate word translations via zero-shot or few-shot prompting. Mapping-based methods are superior to generation-based methods in unsupervised settings, especially are far superior on low-resource languages (Li et al., 2023), primarily due to the unbalanced training corpus size of each language supported by LLMs (Zhu et al., 2023a).

The existing fully unsupervised mapping-based approaches still need to carefully address two issues. First, these approaches rely on the strong assumption that monolingual word embedding spaces are isomorphic and the mapping matrix should be under orthogonal constraint, but this assumption does not hold true for all languages (Søgaard et al., 2018; Glavaš et al., 2019), especially for distant language pairs (Ormazabal et al., 2019; Vulić et al., 2019). Therefore, weak orthogonal constraints have been proposed to tackle this issue (Mohiuddin et al., 2020; Glavaš and Vulić, 2020).

Second, a global mapping matrix does not consistently perform optimally across all subspaces (Nakashole, 2018; Wang et al., 2020). As shown in Figure 1, subspaces exhibit inconsistent structural similarity. With a global orthogonal mapping, BLI accuracy varies among different subspaces: the highest accuracy is 54.3% in subspace 0 and the lowest accuracy is 6.89% in subspace 5. To alleviate the issue, recent research proposed a multi-adversarial learning method (Wang et al., 2020) and a graph-based paradigm (Ren et al., 2020) to learn or refine a specific mapping for each subspace. However, in these approaches, multiple subspaces assigned by initial mappings are static. Once initial solutions of these mappings are not good enough, they may get stuck in poor local optima.

Different from previous methods, we propose a Dynamic Multiple subspaces cross-lingual alignment framework for fully unsupervised Bilingual Lexicon Induction, named DM-BLI. It leverages intra-cluster and inter-cluster contrastive learning to achieve precise alignment at subspace level for both source and target languages, along with dynamically updating the subspace assignment of each word. DM-BLI starts by clustering the embeddings of source language to establish multiple valid subspaces. Then, we induce an initial solution to discover corresponding multiple subspaces in the target language. Finally, we iteratively refine a pair of specific mappings for each subspace pair until convergence is reached.

In summary, we make the following contributions:

- We propose a dynamic multiple subspaces cross-lingual alignment framework for the BLI task, which achieves customized mappings for each subspace pair.
- To boost the performance of our model, we design a contrastive learning framework including intra-cluster and inter-cluster level based on unsupervised clustering to dynamically update the subspace assignment, avoiding falling into local optima.
- We conduct extensive experiments to demonstrate the effectiveness of our method on twelve language pairs including six rich-resource and six low-resource language pairs, and DM-BLI achieves significant improvements especially for distant and low-resource language pairs.

2 Related Work

2.1 Cross-lingual Word Embedding

Bilingual lexicons can be induced via nearest neighbour retrieval on CLWE, which represent lexical words from two or more languages in a shared space.

Based on whether parallel corpora are used or not, CLWE approaches can be categorized into three groups: supervised (Faruqui M, 2014; Zou W Y, 2013; Vulić I, 2015), semi-supervised (Artetxe M, 2017; Patra et al., 2019), and unsupervised approaches (Conneau et al., 2017; Artetxe et al., 2018). Because parallel corpora are not available for many languages, unsupervised approaches gain much more attention.

But unsupervised methods do not require any seed dictionary at all, it is more difficult to induce a reliable initial solution which plays a crucial role in alignment. Therefore, GAN-based adversarial training (Zhang et al., 2017), optimal transport solution (Alvarez-Melis and Jaakkola, 2018), Autoencoder (Mohiuddin and Joty, 2019), and graph-based alignment (Ren et al., 2020) were utilized to better match embedding distribution and find a better initial solution in a fully unsupervised way.

Based on the type of pre-trained monolingual embeddings, CLWE can be divided into two groups: static CLWE and contextual CLWE. Most works focused on static word embeddings (Ruder et al., 2019), which can be derived by Word2Vec (Mikolov et al., 2013b) or fastText (Bojanowski et al., 2016). However, static embeddings lack contextual information to capture polysemy. Therefore, contextual embeddings, generated from monolingual and multilingual pre-trained language models (Devlin et al., 2019; Lample and Conneau, 2019), were utilized as input monolingual embeddings. However, they cannot surpass static embedding in the BLI task based on the same mapping technologies even with much more training time (Vulić et al., 2020; Liu et al., 2021).

2.2 Bilingual Lexicon Induction

Bilingual lexicon induction is the task of inducing word translations from monolingual corpora of two languages.

Existing BLI approaches achieved promising performance on semantically similar and rich-resource language pairs, but were still far from satisfied on distant and low-resource language pairs. For example, unsupervised BLI accuracy on English-Spanish exceeded 80%, while under 40% on English-Chinese (e.g. Conneau et al., 2017; Wang et al., 2020; Ren et al., 2020). In low-resource language pairs like Bulgarian-Hungarian, LLaMA$_{13B}$ achieved 23.61% accuracy, whereas VecMap (Artetxe et al., 2018) achieved 39.24% (Li et al., 2023).

To address this issue, Taitelbaum et al. (2019) suggested leveraging auxiliary languages to bridge the gap between semantically distant and low-resource language pairs. Based on the observation that words are naturally grouped into different semantic subspaces and the BLI accuracies of different subspaces are not uniform, Wang et al. (2020) proposed a multi-adversarial learning method to learn a specific mapping for each subspace. However, this GAN-based method was less robust and its assignment of subspaces was fixed initially which would bring the noise of initial solution.

Different from previous work, we propose a dynamic multiple subspaces alignment framework for unsupervised BLI to achieve more robust and precise alignment at subspace level for both source and target languages, along with dynamically updating the subspace assignment of each word.

3 Methodology

3.1 Formulation

Given the source and target languages, let X and Y be the normalized pre-trained monolingual embeddings for source and target languages, respectively. Our goal is to find the optimal mapping matrices W_X^* and W_Y^*, with which XW_X^* and YW_Y^* are projected in a shared CLWE space, where semantically similar words across languages are close to each other.

Figure 2 illustrates the four procedural steps of our BLI method: multiple subspaces clustering on the source language, initial alignment, intra-cluster and inter-cluster contrastive refinement, and bilingual lexicon induction.

3.2 Multiple Subspaces Discovery

Multiple subspaces discovery contains the first two steps in Figure 2: multiple subspaces clustering and initial alignment. It aims to find pairs of subspaces $\{C_{s_i}, C_{t_i}\}$ from the source and target languages, where $i = 1, 2...K$ and K is the number of subspaces.

Multiple subspaces clustering is only carried on source language embedding to obtain K subspaces. Let $C_{s_i} = \{v_1^{s_i}, v_2^{s_i}, ..., v_n^{s_i}\}$ be the i-th subspace, where $v_k^{s_i}$ is the k-th word in source subspace C_{s_i} and n is the number of words in C_{s_i}. A major challenge in multiple subspaces clustering is to determine the optimal number of subspaces in advance. To tackle this issue, we use a parameter-free hierarchical clustering called First Integer Neighbor Clustering Hierarchy (FINCH) (Sarfraz et al., 2019) to provide a reference number K. Then, K-means algorithm (MacQueen et al., 1967) is used to cluster X into K subspaces.

Then, an initial alignment is conducted for identifying corresponding Ksubspaces in the target language, denoted as $C_{t_i} = \{v_1^{t_i}, v_2^{t_i}, ..., v_m^{t_i}\}$, where $i = 1, 2, \ldots, K$ and $v_j^{t_i}$ is the j-th word in target subspace C_{t_i}. Specifically, we operate the initial alignment following (Artetxe et al., 2018) to get a pair of global initial mapping matrices W_X and W_Y, with which we can retrieve the translation of each target word in the source language. Subsequently, the subspace index of the target word is set to be the subspace index of its translation.

3.3 Multiple Subspaces Contrastive Refinement

A single global mapping does not consistently perform optimally across all subspaces (Nakashole, 2018; Wang et al., 2020). Therefore, the proposed framework will dynamically refine matrices for each subspace pair. This framework contains both inter-cluster and intra-cluster contrastive learning. Inter-cluster contrastive learning ensures the distinguishability of features from different subspaces, thereby facilitating more effective customized mapping. Intra-cluster contrastive learning brings translation pairs within the subspace closer together, while push non-translation pairs further apart, thus achieving finer-grained alignment. The whole refinement process will be completed subspace by subspace.

3.3.1 Inter-cluster Contrastive Learning

Given the subspace pair $\{C_{s_i}, C_{t_i}\}$, inter-cluster contrastive learning aims to bring the whole subspaces C_{s_i} closer to C_{t_i}, while pushing it away from other non-corresponding subspaces $C_{t_j, i \neq j}$.

We introduce optimal transport distance as the metric to evaluate distance of two subspaces distribution, in our work Wasserstein distance (Han et al., 2022) has been applied. The Wasserstein distance between the distributions of two subspaces can be calculated as:

$$\mathbf{D_w}(C_{s_i}, C_{t_i}) = \min_{\boldsymbol{T} \in \pi(C_{s_i}, C_{t_i})} \sum_{j=1}^{n} \sum_{k=1}^{m} \boldsymbol{T}_{jk} \boldsymbol{c}(v_j^{t_i}, v_k^{s_i}) \tag{1}$$

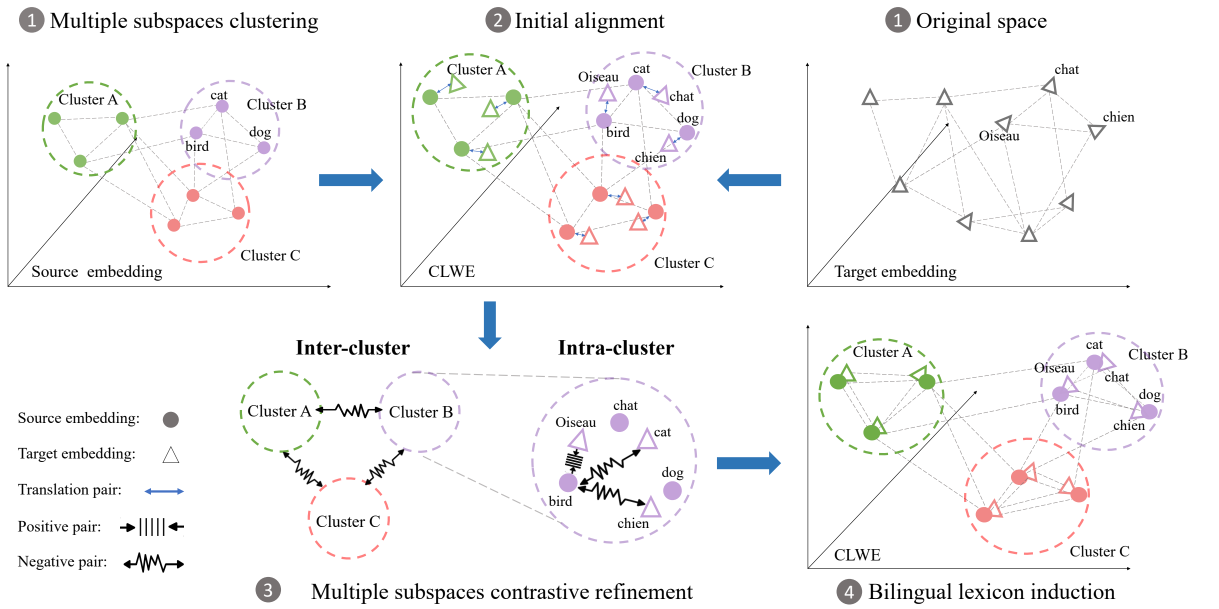

Figure 2. An illustration of the proposed DM-BLI framework. ❶ represents the monolingual word embedding spaces of source and target language, where English is the source language denoted by circles while French is target language denoted by triangles. Multiple subspaces clustering is only applied to source language(English) and different colors represent different subspaces. ❷ represents a cross-lingual word embedding space via an initial alignment. ❸ is a multiple subspaces contrastive learning refinement block aiming to push away words from different clusters and pull closer the words being translation for each other closer within the cluster. ❹ represents refined cross-lingual word embedding space, where words being translation for each other stay closer.

where $\boldsymbol{c}(v_j^{t_i}, v_k^{s_i})$ is the transport cost between words $v_j^{t_i} \in C_{t_i}$ and $v_k^{s_i} \in C_{s_i}$, and $\boldsymbol{T}_{jk}$ represents the transport plan between $v_j^{t_i}$ and $v_k^{s_i}$.

Based on the K pairs of subspaces, we calculate a bi-direction inter-cluster contrastive learning loss as follows:

$$\mathcal{L}_{s2t} = -\frac{1}{K}\Bigg\{\log\left(e^{-\mathbf{D}_w(C_{s_i},C_{t_i})/\tau}\right) + \sum_{j\neq i}\log\left(1 - e^{-\mathbf{D}_w(C_{s_i},C_{t_i})/\tau}\right)\Bigg\}$$

$$\mathcal{L}_{t2s} = -\frac{1}{K}\Bigg\{\log\left(e^{-\mathbf{D}_w(C_{t_i},C_{s_i})/\tau}\right) + \sum_{j\neq i}\log\left(1 - e^{-\mathbf{D}_w(C_{t_i},C_{s_i})/\tau}\right)\Bigg\} \quad (2)$$

where τ is a temperature parameter. To be specific, the aforementioned process is applied to the sampled distribution of subspace, where the proportion of samples is determined by a preset threshold.

Finally, we obtain the final inter-cluster contrastive loss $\mathcal{L}_{inter}$ as below, where λ is the trade-off set to be 0.5 between two directions:

$$\mathcal{L}_{inter} = \lambda * \mathcal{L}_{s2t} + (1-\lambda) * \mathcal{L}_{t2s} \quad (3)$$

3.3.2 Intra-cluster Contrastive Learning

Given the subspace pair $\{C_{s_i}, C_{t_i}\}$, intra-cluster contrastive learning is to ensure word pair $(v_j^{s_i}, v_k^{t_i})$ are closer, which are translations to each other in C_{s_i} and C_{t_i}.

Based on the mapping matrices W_X and W_Y, we can initially construct a bilingual dictionary $\mathbf{D}$ by retrieving the translation of each target word in the source language, where $\mathbf{D} = \{(v_1^{t_i}, v_1^{s_i}), (v_2^{t_i}, v_2^{s_i}), ..., (v_n^{t_i}, v_n^{s_i})\}$ and n is the number of words in $\mathbf{D}$.

However, the quality of $\mathbf{D}$ depends on the quality of mapping matrices. To alleviate the noise brought by the current solution, we selectively sample high-confidence word translation pairs from $\mathbf{D}$, where confidence is determined by the similarity gap between the selected translation and the second candidate translation with the source word.

Based on the sampled translation pairs, the intra-cluster contrastive learning loss can be defined as:

$$\mathcal{L}_{intra} = -\sum_{i=1}^{s}\log\frac{e^{sim(v_i^s, v_i^t)/\tau}}{\sum_{j=1}^{s} e^{sim(v_i^s, v_j^t)/\tau}} \quad (4)$$

Where s is the number of sampled translation pairs and τ is a temperature parameter. Ultimately, the loss of the whole contrastive refinement can be defined as follows:

$$\mathcal{L} = \mathcal{L}_{inter} + \mathcal{L}_{intra} \tag{5}$$

3.4 Multiple Subspaces Dynamic Updating

A single round of subspace assignment may introduce noise from the initial solution, potentially causing CLWEs to fall into local optima. Therefore, we propose to dynamically adjust the subspace assignment of each word in target language during the process of updating W_X and W_Y.

To clarify, the assignment of multiple subspaces in source language $C_s = \{C_{s_1}, C_{s_2}, ..., C_{s_K}\}$ is fixed once the clustering process is completed. For word v_i^t in target language, its translation from source language v_i^s is retrieved based on XW_X and YW_Y. The subspace index of v_i^s will be assigned to v_i^t. Upon updating W_X and W_Y, the subspace assignment of v_i^t will be adjusted accordingly to maintain consistency whenever its translation changes.

As we mentioned before, the whole refinement process will be operated subspace by subspace. For each subspace C_{t_i} in target language, the whole dynamic updating procedure stops until convergence is reached. Convergence can be determined by measuring the overlap of target words within C_{t_i} between the current and previous rounds. Besides, once a subspace has achieved convergence, its assignments are finalized, ensuring that the words within it remain unchanged in their respective subspaces. The whole methodology is summarised in Algorithm 1.

Algorithm 1: Dynamic Multiple Subspaces Alignment for Unsupervised BLI

Input: Monolingual word embedding spaces X, Y
Output: $\{W_{x_i}^*\}_{i=1}^K$, $\{W_{y_i}^*\}_{i=1}^K$

```
 {C_si}_{i=1}^K ← Apply Clustering on X;
 W_X, W_X ← Initial Alignment ;
 {C_ti}_{i=1}^K ← Calculate XW_X, YW_Y;
 for i ≤ K do
     Initialize W_xi, W_yi with W_X, W_Y ;
     while not convergence do
         W_xi, W_yi ← Optimise loss
         C_si ← Keep C_si fixed
         C_ti ← Update C_ti with W_xi, W_yi
 return {W*_xi}_{i=1}^K, {W*_yi}_{i=1}^K;
```

4 Experiment Setup

We evaluate our framework in both supervised and unsupervised BLI tasks on 12 language pairs, which contain 6 rich-resource language pairs: Spanish (ES), German (DE), Russian (RU), Arabic (AR), Japanese (JA) and Chinese (ZH), all cross-lingual to English (EN) and six low-resource language pairs: Finnish (FI), Hindi (HI), Turkish (TR), Indonesian (ID), Bulgarian (BG) and Catalan (CA), all cross-lingual to English (EN).

4.1 Dataset

We use fastText vectors trained on full Wikipedias for each language (Bojanowski et al., 2016) as monolingual word embeddings. We use the widely used MUSE bilingual lexicon (Conneau et al., 2017), released by Facebook, as ground truth lexicon. MUSE provides 110 bilingual lexicons and each lexicon contains the 6,500 most frequently used words in each language, split in a test set of 1,500 words and a training set of 5,000.

4.2 Baselines

Baselines are divided into supervised and unsupervised two lines as described below. We run the released code of each baseline in our experiments.

Supervised BLI

MUSE: Conneau et al. (2017) learned an orthogonal map by minimizing the Euclidean distance between the supervised translation pairs.

VecMap: Artetxe et al. (2018) used a multi-step framework consisting of several steps: whitening, orthogonal mapping, re-weighting, de-whitening, and dimensionality reduction.

BLISS: Patra et al. (2019) proposed a semi-supervised approach with a weak orthogonality constraint in the form of a back-translation loss.

CL-BLI: Li et al. (2023) proposed a robust and effective two-stage contrastive learning framework to combine static and contextual embeddings.

Unsupervised BLI

MUSE: Unsupervised MUSE (Conneau et al., 2017) used adversarial training and iterative Procrustes refinement.

VecMap: Unsupervised VecMap (Artetxe et al., 2018) used intra-linguistic word similarity information to induce initial solution.

Ad. : Mohiuddin and Joty (2019) proposed a adversarial auto-encoder framework, where adversarial mapping was done at the latent embedding space.

Method	Precision@1						Precision@5						Avg.
	FI-$*$	HI-$*$	TR-$*$	ID-$*$	BG-$*$	CA-$*$	FI-$*$	HI-$*$	TR-$*$	ID-$*$	BG-$*$	CA-$*$	
Supervised													
MUSE	46.50	25.65	39.82	35.56	39.28	46.19	66.07	39.17	57.56	50.92	56.62	60.52	46.99
BLISS	49.94	28.17	41.45	38.49	42.21	47.26	68.97	42.43	59.39	54.05	59.51	61.94	49.48
VecMap	58.12	34.07	49.37	44.72	49.13	54.35	75.43	48.40	66.24	59.52	64.62	66.84	55.90
CL-BLI	57.78	32.62	48.52	43.43	47.34	53.89	75.97	47.02	59.93	58.63	64.20	67.09	54.70
DM-BLI	**60.29**	**35.57**	**53.09**	**48.24**	**50.80**	**56.47**	**77.08**	**49.24**	**69.11**	**62.09**	**66.16**	**68.57**	**58.06**
Unsupervised													
MUSE	0.05	0.00	36.82	36.35	38.31	46.07	0.05	0.05	54.76	51.65	55.05	60.51	31.64
VecMap	54.71	28.19	48.92	45.65	45.69	53.52	71.72	41.54	65.25	59.76	61.24	65.63	53.49
Ad.	0.45	0.01	46.69	0.09	0.03	53.06	1.47	0.03	63.08	0.31	0.11	65.55	19.24
BLOOM$_{7B}$	23.43	28.30	30.82	45.45	16.75	43.89	25.75	28.54	34.08	49.77	16.94	48.01	32.64
Llama$_{13B}$	40.98	30.68	44.90	48.63	56.86	48.83	41.64	30.69	45.24	48.95	57.16	49.19	45.31
GPT-3.5	**60.37**	**56.11**	**54.49**	48.37	**67.51**	45.15	64.33	**57.40**	55.99	49.35	**69.53**	45.78	**56.19**
DM-BLI	57.48	30.80	51.98	**48.81**	47.63	**56.15**	**74.10**	43.75	**67.95**	**62.46**	63.36	**67.61**	56.00

Table 1. Precision@1 and Precision@5 for the BLI task on six low-resource language pairs, where $*$ represents EN(English). The best score is shown in **bold**, and the suboptimal score is shown in underlined.

BLOOM$_{7B}$ (Workshop et al., 2022): It is a decoder-only Transformer language model that supports 46 natural languages. 7B parameters version was used in our experiment.

Llama$_{13B}$ (Touvron et al., 2023): It is a decoder-only LLM which supports 20 languages. 13B parameters version was used in our experiment.

GPT-3.5 (Brown et al., 2020): It is a decoder-only LLM with 175B parameters, supported by 38 languages. GPT-3.5-turbo was used in our experiment.

4.3 Implementation details

We choose the most 7,500 frequent vocabularies of each language. The normalization procedure for pre-trained embedding contains three steps: length normalizes the embeddings, then mean centers each dimension, and then length normalizes them again.

For multiple subspaces discovery, the number of subspaces is set to be 9 and we will discuss the impact of this setting later. For inter-cluster contrastive learning, only words with weight above 0.45 are sampled to represent the subspace distribution. For intra-cluster contrastive learning, we only sample the top 20% of word translation pairs sorted descending by confidence.

Following the previous research (Patra et al., 2019), the prompt template for **Llama$_{13B}$** is defined as: "Translate from L^x to L_y: w^x=>"; the prompt template for **GPT-3.5** is defined as: "Translate the L^x word w^x into L_y:". Both of them are provided as the best template for each of them in Li et al. (2023).

The evaluation for BLI is done by comparing the bilingual lexicon constructed by each model with the benchmark lexicon MUSE (Conneau et al., 2017) and reporting precision Precision@N for $\mathbf{N} = 1, 5$. Precision@N accounts for accuracy for which the correct translation of the source words is in the N-th nearest neighbors based on CSLS (Conneau et al., 2017).

5 Result and Discussion

5.1 Results in Low-resource Languages

Table 1 summarizes the results of the supervised and unsupervised BLI tasks in low-resource language pairs. In both tasks, our proposed method shows significant improvements, particularly in Precision@5, with an average of 2.16 points higher than the strongest baseline VecMap in the supervised task. In the unsupervised task, our method performs nearly as well as the strong baseline GPT-3.5.

In the supervised task, DM-BLI outperforms all the baseline methods on all language pairs, demonstrating the robustness and effectiveness of our framework on low-resource language pairs. In the unsupervised task, DM-BLI outperforms all the baseline methods on four out of six language pairs and archives suboptimal scores in the remaining pairs at Precision@5. It demonstrates that our method is competitive even compared with GPT-3.5, which has 175B parameters and supports 38 languages. The unsatisfied performance

Method	Precision@1						Precision@5						Avg.
	ES-$*$	DE-$*$	RU-$*$	AR-$*$	JA-$*$	ZH-$*$	ES-$*$	DE-$*$	RU-$*$	AR-$*$	JA-$*$	ZH-$*$	
Supervised													
MUSE	67.80	63.14	53.23	44.33	0.14	8.29	78.13	75.86	70.19	61.16	0.41	18.87	45.13
BLISS	68.46	63.49	54.88	45.70	0.01	6.43	78.86	76.69	71.28	62.47	0.04	14.00	45.19
VecMap	71.70	66.46	59.58	51.54	37.14	42.50	80.43	78.22	74.69	67.00	53.65	62.23	62.10
CL-BLI	**73.02**	**69.00**	61.31	**53.14**	35.07	42.44	**81.71**	**80.28**	**77.10**	**68.95**	50.68	62.26	62.91
DM-BLI	72.87	68.28	**61.61**	52.33	**41.03**	**44.83**	81.16	79.35	76.35	67.80	**56.94**	**64.13**	**63.89**
Unsupervised													
MUSE	67.89	63.27	50.49	0.03	0.09	0.01	78.37	75.87	67.10	0.08	0.37	0.04	33.63
VecMap	72.00	67.17	56.42	47.43	26.62	33.39	79.91	77.77	71.45	63.53	40.62	51.86	57.35
Ad.	71.93	66.63	55.50	0.00	0.00	0.00	79.99	77.59	70.56	0.00	0.01	0.01	35.19
BLOOM$_{7B}$	52.50	38.34	26.06	32.67	21.34	34.35	56.19	41.49	26.27	32.80	21.38	34.53	34.83
Llama$_{13B}$	60.58	57.80	64.44	22.13	38.56	32.28	61.09	58.51	65.10	22.14	38.57	32.29	46.12
GPT-3.5	68.17	63.07	**74.15**	**65.94**	**71.80**	**65.12**	70.72	66.08	**76.84**	**69.88**	**74.95**	**68.69**	**69.62**
DM-BLI	**72.94**	**68.67**	58.91	48.58	32.42	37.34	**80.65**	**78.92**	73.45	64.70	47.98	56.45	60.08

Table 2. Precision@1 and Precision@5 for the BLI task on six rich-resource language pairs, where $*$ represents EN(English). The best score for is shown in **bold**, and the suboptimal score is shown in underlined.

of BLOOM$_{7B}$ and Llama$_{13B}$ also suggests that the generalization of LLMs to low-resource languages remains an open challenge.

5.2 Results in Rich-resource Languages

Table 2 summarizes the main results of the supervised and the unsupervised BLI tasks on rich-resource language pairs.

In supervised tasks, our proposed method achieves significant improvements, with average nearly 1 point higher than the strongest baseline CL-BLI. We achieve the optimal or sub-optimal performance on all the language pairs. Notably, our method achieves a 6.26% improvement over CL-BLI on distant language pairs Japanese to English, demonstrating advantages of multiple subspace alignment on distant language pairs.

In unsupervised tasks, DM-BLI achieves the sub-optimal result on rich-resource language pairs. While it outperforms the previous mapping-based SOTA method VecMap but underperforms GPT-3.5. The outstanding performance of GPT-3.5 verifies the potential of the latest generation of LLMs for developing bilingual lexicons with sufficient training and a large amount of parameters. However, BLOOM$_{7B}$ and Llama$_{13B}$ are still far lagging behind the traditional mapping-based method even on rich-resource language pairs, which verifies that it is difficult to extract lexical information from large language models (Liu et al., 2021).

5.3 Influence of Translation Direction

In this subsection, we examine how the translation direction affects BLI results in unsupervised setup. The language pairs we choose as examples are Japanese (JA), Chinese (ZH), Finish (FI), Indonesian (ID) from and to English (EN), as shown in Table 3.

Method	EN-JA		EN-ZH		EN-FI		EN-TR	
	→	←	→	←	→	←	→	←
MUSE	0.01	0.37	0.01	0.04	0.06	0.05	30.73	54.76
VecMap	35.63	40.62	32.62	56.45	43.08	71.72	40.10	65.29
GPT-3.5	**57.06**	**74.98**	**42.56**	**68.69**	**58.97**	64.33	**52.63**	55.99
DM-BLI	39.43	47.98	34.69	56.45	44.30	**74.10**	41.90	**67.95**

Table 3. Precision@5 for the bi-direction unsupervised BLI task on four language pairs. The best score is shown in **bold**, the suboptimal score is shown in underlined.

From Table 3, we observe the performance differences in the two directions of the language pair. Specifically, the results from English to other languages significantly lag behind those from other languages to English. A part of the reason is that there are more unique English words than non-English words in the evaluation set (Xu et al., 2018). It also proves that LLMs exhibit unbalanced capacities across languages, performing better at translating into English than translating into non-English (Zhu et al., 2023b).

5.4 Influence of the Number of Subspaces

In this section, we discuss the impact of the number of subspaces on performance of DM-BLI, taking distant language pair JA2EN as an example.

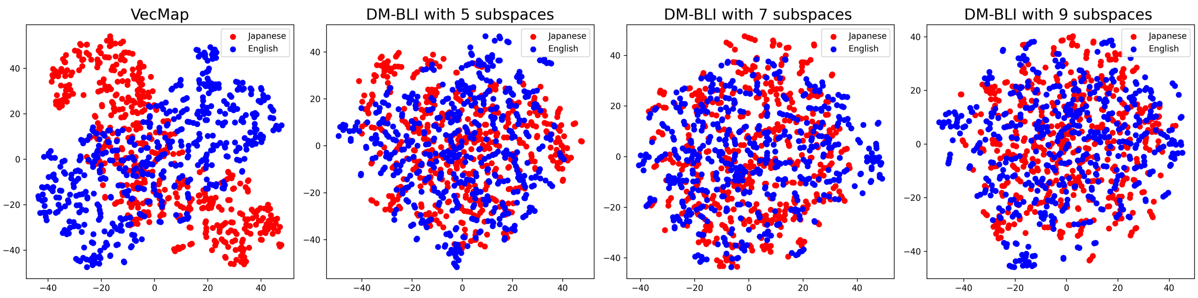

Figure 3. t-SNE visualization of sampled CLWEs derived from VecMap and DM-BLI, where visualization of CLWE derived from DM-BLI is based on different numbers of multiple subspaces.

As shown in Figure 3, compared with VecMap who only use a global mapping, our method lets word with same meaning from different languages get much closer in a shared CLWEs space via multiple subspace-level alignments.

Notably, from Figure 3, we can find that even using different numbers of subspaces, DM-BLI still achieved nearly the same results, which shows that it is not sensitive to the number of subspaces and further proves the robustness of our method.

5.5 Effect of Multiple Subspaces Alignment

Notice that our method focuses on leveraging multiple subspace alignments to achieve better performance for BLI. In this subsection, we discuss the advantages of multiple subspaces alignment from our method DM-BLI, taking low-resource language pair CA2EN as an example.

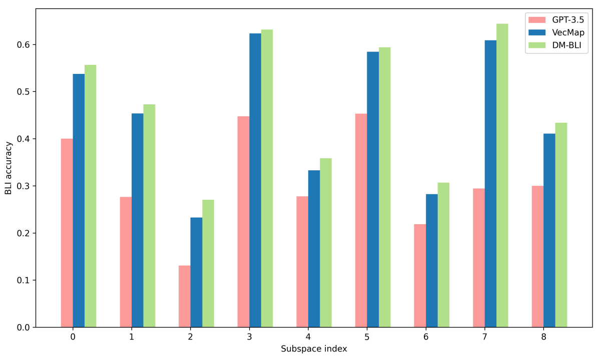

Figure 4. Precision@1 for unsupervised BLI from Catalan to English in different English subspaces.

As shown in Figure 4, on low-resource language pair like CA2EN, we can find that BLI accuracies for all subspaces based on DM-BLI are higher than the strongest mapping-based baseline VecMap. Notably, we also find that unbalanced alignments occur in a generative way via GPT-3.5 as well. Furthermore, LLM's capability on BLI is still far lagging behind mapping-based approach.

In order to show effect of DM-BLI more intuitively, we sample 2 subspaces for visualization. As shown in Figure 5, via multiple subspaces alignment, translation pairs within the subspace stay closer together than applying a global mapping.

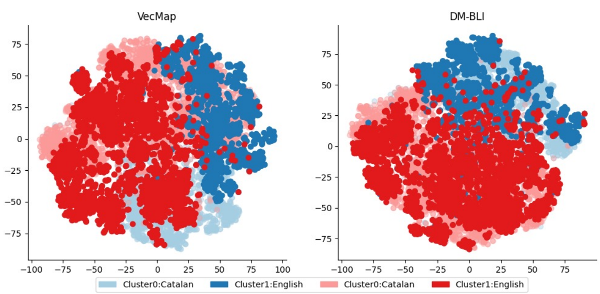

Figure 5. t-SNE visualization of two sampled subspaces in CLWE space derived from VecMap and DM-BLI on CA2EN. Within the subspace, dots denoted by the same color but different transparency are translation pairs.

6 Conclusion

In this paper, we propose a Dynamic Multiple subspaces alignment framework for unsupervised BLI, called DM-BLI. Our method utilizes multiple subspaces alignment instead of a single mapping alignment to achieve more accurate alignment on the subspace level. The experiments show that our method can significantly improve the bilingual word induction performance compared with strong baselines even including GPT-3.5, especially for distant and low-resource language pairs. At the same time, the unsatisfied performances of BLOOM_{7B} and Llama_{13B} on all language pairs also suggest that it is difficult to extract lexical information from large language models and the generalization of LLMs to low-resource languages remains an open challenge. In the future, we will consider combining our method with multilingual LLMs to take advantage of these two paradigms.

Limitations

First, due to our limited computing resources, we did not conduct a comprehensive evaluation of the BLI capabilities of multilingual LLMs. For open-source LLMs, LLMs exceeding 13B parameters were not evaluated in the experiment. For close-source LLMs, experiments were mainly conducted on GPT-3.5-turbo which is not the latest and best.

Second, public BLI datasets are not enough to support a comprehensive evaluation. In the evaluation standard dictionary, the proportion of ground-truth translations in different categories is uneven. As also discussed in (Li et al., 2023), current evaluation will not work for words that are not included in the gold translations.

Acknowledgements

References

David Alvarez-Melis and T. Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Conference on Empirical Methods in Natural Language Processing*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. 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, Melbourne, Australia. Association for Computational Linguistics.

Agirre E Artetxe M, Labaka G. 2017. Learning bilingual word embeddings with (almost) no bilingual data. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability. *arXiv preprint arXiv:2305.10266*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tobias Eder, Viktor Hangya, and Alexander Fraser. 2021. Anchor-based bilingual word embeddings for low-resource languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 227–232, Online. Association for Computational Linguistics.

Dyer C Faruqui M. 2014. Improving vector space word representations using multilingual correlation. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics*.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing word-level translations from multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56, Online. Association for Computational Linguistics.

Yuehui Han, Le Hui, Haobo Jiang, Jianjun Qian, and Jin Xie. 2022. Generative subgraph contrast for self-supervised graph representation learning. In *European Conference on Computer Vision*, pages 91–107. Springer.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2023. On bilingual lexicon induction with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9577–9599, Singapore. Association for Computational Linguistics.

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2022. Improving word translation via two-stage contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4353–4374, Dublin, Ireland. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, Carey Priebe, and Philipp Koehn. 2022. Bilingual lexicon induction for low-resource languages using graph matching via optimal transport. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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, Online. Association for Computational Linguistics.

Tasnim Mohiuddin and Shafiq Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3857–3867, Minneapolis, Minnesota.

Ndapa Nakashole. 2018. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522, Brussels, Belgium. Association for Computational Linguistics.

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 the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Shuo Ren, Shujie Liu, Ming Zhou, and Shuai Ma. 2020. A graph-based coarse-to-fine method for unsupervised bilingual lexicon induction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3476–3485, Online. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. 2019. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8934–8943.

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, Melbourne, Australia. Association for Computational Linguistics.

Hagai Taitelbaum, Gal Chechik, and Jacob Goldberger. 2019. Multilingual word translation using auxiliary languages. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1330–1335, Hong Kong, China.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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 (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Moens M F Vulić I. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bi-lingual lexicon induction. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Beijing, China.*

Haozhou Wang, James Henderson, and Paola Merlo. 2020. Multi-adversarial learning for cross-lingual word embeddings. *arXiv preprint arXiv:2010.08432*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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, Brussels, Belgium. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023a. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023b. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Cer D et al Zou W Y, Socher R. 2013. Bilingual word embeddings for phrase-based machine translation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics*.