基于动态子空间重构的跨语言词向量对齐及应用

顾晓洋，胡玲，徐月梅
北京外国语大学信息科学技术学院，北京，100089
gxy@bfsu.edu.cn,huling@bfsu.edu.cn,xuyuemei@bfsu.edu.cn

摘要

无监督双语词典归纳（Bilingual Lexicon Induction，BLI）通过学习映射函数对齐两种不同语言的单语词嵌入空间，从而推导单词翻译，在相似语言对中取得显著成功。然而，传统方法依赖单一线性映射，在远距离或低资源语言对上性能欠佳。为解决此问题，本文提出DM-BLI，一个基于动态多子空间对齐的无监督双语词典归纳算法及其应用框架。首先，DM-BLI通过多子空间映射提升对齐精度，重构源语言词嵌入空间，采用无监督聚类识别子空间，结合粗略全局对齐定位目标空间对应子空间，并通过簇内和簇间对比学习优化映射矩阵。在包含5个高资源和5个低资源语言对的有监督和无监督实验中显著提升性能。此外，DM-BLI基于所构建的词典使用logits lens技术评估大语言模型（Large Language Model, LLM）的跨语言能力，通过翻译和重复任务计算余弦相似度，结合词向量空间语义特征验证模型生成翻译的语义合理性。相较传统LLM的跨语言评估方法仅以静态的BLI翻译对为标准，DM-BLI能识别未被词典覆盖但语义合理的翻译，显著提升评估的鲁棒性和语义泛化能力，更准确全面地衡量大语言模型的跨语言语义映射能力。我们的代码发布https://github.com/huling-2/DM-BLI.git.

关键词： 跨语言词嵌入；词嵌入对齐；对比学习；无监督聚类

Dynamic Multi-Subspace Alignment for Cross-Lingual Word Embeddings and Its Applications

Xiaoyang Gu, Ling Hu, Yuemei Xu
School of Information Science and Technology,
Beijing Foreign Studies University, Beijing 100089, China
gxy@bfsu.edu.cn,huling@bfsu.edu.cn,xuyuemei@bfsu.edu.cn

Abstract

Unsupervised bilingual lexicon induction (BLI) aligns monolingual word embedding spaces to infer word translations, achieving success in similar language pairs. However, reliance on a single linear mapping leads to poor performance for distant or low-resource language pairs. We introduce DM-BLI, a dynamic multi-subspace alignment framework for unsupervised BLI. DM-BLI enhances alignment by using multiple subspace mappings, identifying subspaces via unsupervised clustering in the source embedding space,

基金项目：国家社科基金项目“语言安全视域下大语言模型的价值观生成机制和风险应对研究”（No.24CYY107）；中央高校基本科研业务费专项“生成式人工智能的价值观对齐研究”（No.2024TD001）

and locating corresponding target subspaces with coarse global alignment. Contrastive learning refines subspace mappings. Experiments on 10 language pairs show substantial gains. Additionally, we use the DM-BLI lexicon to evaluate large language models' cross-lingual capabilities through translation and repetition tasks, computing cosine similarity and validating semantic reasonableness with embedding space features. Unlike traditional BLI, DM-BLI identifies semantically valid translations not in the lexicon, enhancing evaluation robustness and semantic generalization. We release our code at https://github.com/huling-2/DM-BLI.git.

Keywords: Cross-Lingual Word Embedding , Word Embedding Alignment , Contrastive Learning , Unsupervised Clustering

1 引言

双语词典归纳（Bilingual Dictionary Induction, BLI）旨在寻找源语言单词在目标语言中的翻译，能有效解决大部分非英语语言面临的平行语料资源匮乏问题，实现不同语言之间的知识迁移(Eder et al., 2021; Marchisio et al., 2022a)。

现有的双语词典归纳方法大致可分为两类：基于映射的方法(Conneau et al., 2018; Artetxe et al., 2018; Ren et al., 2020; Li et al., 2022)和基于生成的方法(Gonen et al., 2020; Zhang et al., 2023a; Li et al., 2023)。基于映射的方法旨在通过线性或非线性映射将来自不同语言的单语词嵌入映射至同一个跨语言词向量（Cross-Lingual Word Embedding，CLWE）空间中，并且在此空间中不同语言间语义相近的词尽可能接近。而基于生成的方法则利用大型语言模型（Large Language Models, LLMs）的机器翻译能力(Briakou et al., 2023)，通过零样本或少样本提示直接生成单词翻译。在无监督的情况下，基于映射的方法优于基于生成的方法，特别是在低资源语言上(Li et al., 2023)。这是由于LLMs每种语言的训练语料库大小不平衡，进而导致其在不同语言上存在能力差异(Zhu et al., 2024)。

尽管关于无监督双语词典归纳的研究已经取得一定的进展，但这一任务仍然面临着重要的挑战。首先，大多数基于映射的方法都建立在所有单语词嵌入空间同构的强假设上，并且通常约束映射矩阵为正交矩阵。但是这个强假设在很多情况下并不成立，特别是对于形态学和词源学差异大的远距离语言对(比如，英语-中文)(Søgaard et al., 2018; Glavaš et al., 2019)。因此，弱正交约束被应用于解决这个问题(Mohiuddin et al., 2020a; Glavaš and Vulić, 2020)；第二，在映射的过程中，现有的方法通常只使用一个或一对全局映射将两组单语词嵌入编码至同一语义空间中。但是全局映射聚焦于全局特征，忽略了局部特征，不能在所有子空间上效果达成一致(Nakashole, 2018; Wang et al., 2021)。如图1 所示，不同颜色代表不同的子空间，英语-日语的子空间对之间存在明显不一致的结构相似性。通过单一的全局正交映射，6个子空间对的BLI（词汇翻译）准确率分别为54.3%、48.7%、40.1%、19.4%、18.9%和6.9%。

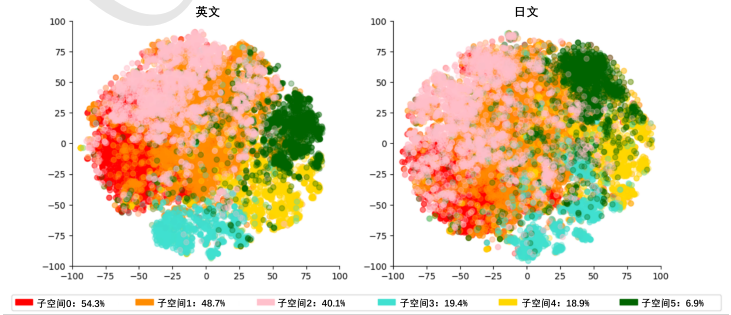


图 1: 英语（左）和日语（右）的单语词嵌入聚类结果t-SNE可视化

为了解决子空间效果不平衡的问题，有研究提出基于多子空间对抗学习的办法(Wang et al., 2021)和基于图的学习范式(Ren et al., 2020)。然而，在这些方法中，单语词向量初始化的多子空间分配是固定不变的，这使得BLI的最终表现强依赖于初始解和初始子空间的质量。一旦所获取的初始解对齐性能差，就会导致结果陷入局部最优。

不同于现有的方法，本文针对于无监督的双语词典归纳任务提出一种名为DM-BLI的动态多子空间跨语言对齐框架。该方法利用簇类和簇间对比学习，实现了对源语言和目标语言的子空间级的精确对齐，并动态更新每个词的子空间分配，避免算法陷入局部最优。DM-BLI首先通过对源语言嵌入进行重构和聚类，建立多个有效子空间。随后，基于初始解发现目标语言中对应的多个子空间。最后，我们迭代优化每个子空间对的特定映射，直至收敛。与之前的方法相比，DM-BLI在无监督和有监督的双语词典归纳任务上效果都更加优异。

总结来说，本文的主要贡献如下：1）提出了一种面向双语词典归纳任务的动态多子空间跨语言对齐框架，为每对子空间提供了定制化映射，实现全局和局部两个维度上更精确的对齐。2）设计了基于无监督聚类的簇内和簇间对比学习框架，动态更新子空间分配，避免算法陷入局部最优。3）进行了大量的实验验证，在10种语言对包括5对高资源和5对低资源语言对上展示了该方法的有效性，特别是在远距离、低资源语言对场景中，DM-BLI相较于VecMap等基线方法均有提升。

2 相关工作

2.1 跨语言词向量

跨语言词向量将不同语言的单词表示在同一空间中，使得不同语言中语义相似的词尽可能接近，双语词典可以通过对其进行近邻检索得到。通过跨语言词向量表示，可将在资源丰富的源语言上训练的模型或知识，迁移到资源匮乏的目标语言上。根据是否使用平行语料，现有的跨语言词向量方法可以分为有监督、半监督和无监督三类。

早期的跨语言词向量嵌入模型主要基于有监督方法，通过大量人工标注的数据，例如单词级别(Faruqui and Dyer, 2014)、句子级别(Zou et al., 2013)和文档级别(Vuli and Moens, 2015)的平行语料库，来对齐源语言和目标语言的单语词嵌入。然而，获取这样的高质量标注数据对大多数语言来说具有挑战性。因此，半监督方法(Agirre et al., 2017; Patra et al., 2019)被提出，它尝试用更小规模的语料或者种子词典减少对跨语言监督信息的依赖。其核心是从初始的少量种子字典开始，然后逐步训练迭代扩展种子字典规模，初始种子词典的规模可减至最小25对(Agirre et al., 2017)。

无监督的跨语言词向量方法首次在2015年被提出(Gouws et al., 2015)，并逐渐成为跨语言词向量嵌入领域的主流方法(Conneau et al., 2018; Artetxe et al., 2018)，主要是因为这些方法不需要任何平行语料库或种子词典，而且更适用于更广泛的语言。与有监督和半监督的方法相比，无监督的方法在获取可靠的初始解这个问题上面临更多的挑战。借助大规模的非平行语料资源，生成对抗网络（GAN）的对抗性训练(Zhang et al., 2017)、最优输运解决方案(Alvarez-Melis and Jaakkola, 2018)、自编码器(Mohiuddin and Joty, 2019)和基于图的对齐(Ren et al., 2020)等方法被用来挖掘两种语言表示之间存在的关系，获取可靠的初始解并进行迭代。

虽然无监督的跨语言词向量生成减少了对双语平行语料或者双语词典的依赖，在性能上也有较好的表现，但仍存在一定的缺点。Søgaard等人(Søgaard et al., 2018)研究发现，基于无监督的CLWE生成对于语言对的选择非常敏感。对于远距离语言对（例如英语-日语），依靠完全无监督的CLWE难以得到高质量的双语词向量词典(Glavaš and Vulić, 2020)。大多数CLWE方法都是通过学习一个正交映射实现，正交约束依赖于单语空间同构的假设。然而当两种语言在词源学和类型学（etymologically and typologically）上相距甚远时，并不满足同构假设，正交映射会失效。为此，Goran等人(Glavaš and Vulić, 2020)放弃学习线性投影矩阵，转而提出一个非参数模型INSTAMAP，通过迭代来学习非线性投影。Marchisio等人(Marchisio et al., 2022b)提出在训练单语词向量的时候加入同构约束，使得训练得到的单语空间满足同构条件。

根据单语词嵌入的种类，现有的跨语言词嵌入方法可以分为静态词嵌入和上下文词嵌入两类。大多数研究集中在静态词嵌入上(Ruder et al., 2019)，这些词嵌入可以由Word2Vec(Mikolov et al., 2013)或fastText(Bojanowski et al., 2016)等工具生成。然而，静态词嵌入缺乏上下文信息，难以捕捉一词多义的特性。因此，有研究开始从单语和多语预训练语言模型(Devlin et al., 2019; Conneau and Lample, 2019)中生成上下文词嵌入，并将其作为单词词嵌入。然而，即使在更长的训练时间下，它们在相同的映射技术下也无法超越静态嵌入在双语词典归纳任务中的表现(Vulić et al., 2020; Liu et al., 2021)。

2.2 双语词典归纳

双语词典归纳旨在从两种不同语言的单语语料库中推导出单词的翻译。模型BLI的性能受语言差异的影响显著，不同语言对之间的变化显著。

BLI方法在语义上相似且资源丰富的语言对上表现良好，但在语义距离较远或资源稀缺的语言对上表现较差。例如，无监督BLI在英语-西班牙语上的准确率超过80%，而在英语-中文上不到40%(Conneau et al., 2018; Ren et al., 2020; Wang et al., 2021)。因此，越来越多的研究关注于如何解决语义距离较远或资源稀缺语言对带来的挑战。

然而，并没有统一的标准来定义资源稀缺的语言对。例如，Zhang等人(Zhang et al., 2023b)使用Twitter上的语言频率作为标准，而Goyal等人(Goyal et al., 2022)根据它们与英语的平行语料资源的数量将语言分类为四类。与英语、西班牙语和中文等高资源语言相比，许多语言尽管它们拥有相当数量的资源但仍然面临着类似极低资源语言的挑战。此外，由于低资源语料少，训练得到的词嵌入质量较差，CLWE通常在真正低资源语言中表现不佳(Michel et al., 2020)。因此，大多数以前的研究集中在相对低资源的语言（如芬兰语和印地语）上，而不是绝对的低资源语言(Mohiuddin et al., 2020b; Tian et al., 2022)。

为了应对远距离和低资源语言对的挑战，Taitelbaum等人(Taitelbaum et al., 2019)建议利用辅助语言来补足语义距离较远和低资源语言对之间的差距。基于观察到单词自然地分组为不同的语义子空间，并且不同子空间的BLI准确性不均匀，Wang等人(Wang et al., 2021)提出了一种多子空间对抗学习的方法，为每个子空间学习特定的映射。然而，这种基于GAN的方法鲁棒性较差，且其子空间分配最初是固定的，可能会引入初始解的噪音。

与以往的研究不同，我们提出了一种动态多子空间对齐的无监督BLI框架，在子空间级别为源语言和目标语言实现更强大和精确的对齐，同时动态更新每个单词的子空间分配，避免算法陷入局部最优。

3 DM-BLI方法概述

3.1 问题定义

基于给定的源语言和目标语言，我们分别令$\boldsymbol{X} \in R^{N\times d}$ 和$\boldsymbol{Y} \in R^{M\times d}$为标准化的预训练单语词向量，其中$M$ 和N 分别表示源语言和目标语言的单词数量，d 表示向量维度。我们的目标是找到最佳映射矩阵$\mathbf{W}_X$ 和$\mathbf{W}_Y$，使得映射后的词向量$\mathbf{W}_X\boldsymbol{X}$ 和$\mathbf{W}_Y\boldsymbol{Y}$位于一个共享的空间中。双语词典归纳可以通过对跨语言词向量进行检索实现。

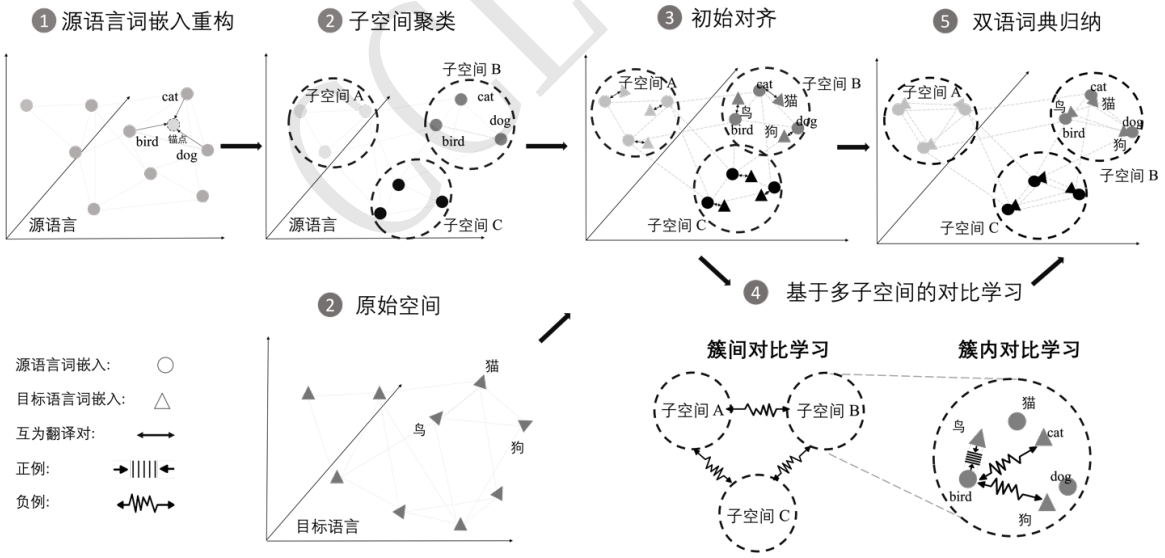


图 2: DM-BLI流程示意图，其中◯代表英语，△代表中文

以英语-中文为例，本文提出的DM-BLI模型框架如图2 所示。DM-BLI共包括五个步骤：源语言词嵌入的重构、源语言的多子空间聚类、初始对齐、基于对比学习的簇内和簇间微调以及双语词典归纳的实现。

3.2 多子空间发现

多子空间发现的目标是发现源语言和目标语言对应的子空间对$\{X_i, Y_i\}$，其中X_i中的单词与Y_i中的单词语义相近，$i = 1, 2, \ldots, K$，K是子空间的数量。多子空间发现的流程主要包含两个步骤：源语言的多子空间聚类以及初始对齐。

3.2.1 源语言的多子空间聚类

首先，我们仅对源语言预训练词嵌入X进行聚类以获得K个有效的子空间，表示为$X = \{X_1, X_2, \ldots, X_K\}$。其中，$X_i \in R^{N_i \times d}$表示第$i$个子空间，$N_i$是其包含的单词数量。然而，对词嵌入空间进行聚类时，面临两个主要挑战：第一，单个词嵌入提供的信息有限，直接对预训练的单词语嵌入进行聚类忽略了语义空间中的结构信息并且会引入原始预训练词向量携带的噪音；第二，如何预先确定子空间的最佳数量，主流的聚类方法如K-means(MacQueen and others, 1967)等受超参数影响大，子空间数量这个变量可能会对后续的结果产生扰动，如何确定一个有效的子空间数量是一个至关重要的问题。

为了解决第一个挑战，我们在对词嵌入空间进行聚类前引入一个类GNNs模块对词嵌入进行预处理。具体地说，通过将整个词嵌入空间视为一个图，通过挖掘图结构的拓扑关联来实现更深层次的语义关联探索，并对词嵌入空间进行处理，使得语义相近的词离得更近，使得后续聚类的置信度更高。对于X中的任一个单词w_x，我们首先将其视为一个节点，与其最邻近的n个单词$w_x^1, w_x^2, \ldots, w_x^n$则是与之相连的$n$个节点，其中每条相连边的权重为两个节点的余弦相似度。同时，为了尽可能地引入噪声，我们对w_x所有相连边的权重做softmax处理。其次，对于任一个单词w_x，我们可以计算获得一个锚点$\widehat{w}_x$，具体值为w_x与最邻近的n个词嵌入的平均值。如公式（1）所示，我们通过计算最邻近的n个单词与锚点$\widehat{w}_x$的差异对w_x进行改造，使得w_x与其语言相近的词空间距离更近，其中p是控制重构比率的超参数。

$$w_x = w_x + p * \sum_{i=1}^{n} weight_i(w_x^i - \widehat{w}_x) \tag{1}$$

为了解决第二个挑战，我们使用一种无参数的层次聚类方法FINCH（First Integer Neighbor Clus-tering Hierarchy）(Sarfraz et al., 2019)，来提供参考子空间参考数量K。然后，使用K-means算法(MacQueen and others, 1967)将X聚类为K个子空间。

3.2.2 初始对齐

基于源语言的聚类结果，初始对齐的目标是在目标语言Y中识别出与源语言$X = \{X_1, X_2, \ldots, X_K\}$相对应的$K$个子空间$Y = \{Y_1, Y_2, \ldots, Y_K\}$，从而实现多子空间对的配对。其中，$Y_i \in R^{M_i \times d}$表示第$i$个子空间，$M_i$是其包含的单词数量。具体来说，我们遵循VecMap(Artetxe et al., 2018)的范式进行初始对齐以获得一对全局初始映射矩阵$W_{X_{init}}$和$W_{Y_{init}}$。通过这些矩阵我们可以检索每个目标单词在源语言中的翻译，其翻译的子空间索引将被分配给对应的目标单词。

3.3 多子空间对比微调

单一的全局映射在所有子空间对上并不总是表现最佳(Nakashole, 2018; Wang et al., 2021)。为了实现更细粒度的对齐，我们提出了一种新颖的多子空间对比学习框架，该算法动态优化每个子空间对的映射矩阵，从而提高BLI性能。这个框架包含簇间和簇内的对比学习两个部分。簇间对比学习的目的是确保不同子空间对之间特征的可区分性，从而更有效地针对每个子空间对进行个性化映射；簇内对比学习的目的是使得子空间对内的源语言和目标语言语义相似的词对距离更近，从而实现更细粒度的对齐。整个优化过程将逐个子空间对完成。

3.3.1 簇间对比学习

给定子空间对$\{X_i, Y_i\}_{i=1}^{K}$，簇间对比学习的目的是强化不同子空间对之间差异性，提高每个子空间对的辨识度。具体地说，簇间对比学习使得已配对的子空间$\{X_i, Y_i\}$离得近，非配对的两个子空间$\{X_i, Y_{j \neq i}\}$离得远。

我们引入了最佳传输距离(本文中为Wasserstein距离(Han et al., 2022))作为评估两个子空间分布距离的度量。与简单的距离度量（如欧几里得距离）相比，Wasserstein距离考虑了概率

分布的整体结构，使其对离群点具有鲁棒性并能更有效地捕捉几何细节。两个子空间分布之间的Wasserstein距离可以计算为：

$$\boldsymbol{D}_W(X_i, Y_i) = \min \sum_{j=1}^{N_i} \sum_{k=1}^{M_i} T_{jk} \boldsymbol{c}\left(w_j^{X_i}, w_k^{Y_i}\right) \tag{2}$$

其中$\boldsymbol{c}\left(w_j^{X_i}, w_k^{Y_i}\right)$代表第$i$个源语言子空间的第$j$个单词$w_j^{X_i}$和第$i$个目标语言子空间的第$k$个单词$w_k^{Y_i}$之间的传输成本，$T_{jk}$表示单词$w_j^{X_i}$和$w_k^{Y_i}$之间的传输计划。

基于给定的K个子空间对，我们设计了一个双向的簇间对比学习损失，如下所示：

$$\mathcal{L}_{s2t} = -\frac{1}{K}\left\{\begin{array}{l}\log\left(e^{-\frac{d_W(X_i,Y_i)}{\tau}}\right) \\ +\sum_{j\neq i}\left(1-e^{-\frac{d_W(X_i,Y_j)}{\tau}}\right)\end{array}\right. \tag{3}$$

$$\mathcal{L}_{t2s} = -\frac{1}{K}\left\{\begin{array}{l}\log\left(e^{-\frac{d_W(Y_i,X_i)}{\tau}}\right) \\ +\sum_{j\neq i}\left(1-e^{-\frac{d_W(Y_i,X_j)}{\tau}}\right)\end{array}\right. \tag{4}$$

其中τ是一个温度参数，$s2t$表示源到目标的相似性得分，反映X_i 在目标空间中的对齐质量，$t2s$表示目标到源的相似性得分，反映Y_j 在源空间中的对齐质量，与$s2t$ 对称，确保双向对齐的平衡。为避免计算量和计算复杂度过大，我们根据单词的重要程度对每个子空间的分布进行采样，采样的比例为$Sample_\alpha$。样本的重要程度由其与相近单词的相似度总和计算获得。每个样本的相近单词数量由预设的阈值决定。

最终的簇间对比损失$\mathcal{L}_{inter}$如公式（5）所示，其中λ设置为0.5：

$$\mathcal{L}_{inter} = \lambda * \mathcal{L}_{s2t} + (1-\lambda) * \mathcal{L}_{t2s} \tag{5}$$

3.3.2 簇内对比学习

给定子空间对$\{X_i, Y_i\}_{i=1}^K$，簇内对比学习的目的是确保在此子空间对内，互为翻译对的单词对$(w_j^{X_i}, w_k^{Y_i})$离得更近。

基于映射矩阵$\mathbf{W}_X$和$\mathbf{W}_Y$，我们可以通过检索每个目标单词在源语言中的翻译构建一个双语词典$\mathcal{D}$。其中，$\mathcal{D} = \{(w_1^{X_i}, w_1^{Y_i}), (w_2^{X_i}, w_2^{Y_i}), \ldots, (w_l^{X_i}, w_l^{Y_i})\}$，$l$是词典中翻译对的数量。

然而，双语词典$\mathcal{D}$的质量依赖于映射矩阵的质量。为了减少引入当前映射带来的噪音，我们会采样置信度较高的翻译对纳入最终的双语词典$\mathcal{D}_s$中。每对翻译的置信度取决于所选翻译与第二候选翻译在源语言单词上的相似度差异。

基于采样得到的高质量翻译对$\mathcal{D}_s$，簇内对比学习损失设计如下：

$$\mathcal{L}_{intra} = -\sum_{i=1}^{|\mathcal{D}_s|} \log \frac{e^{sim(w_i^{X_i}, w_i^{Y_i})/\tau}}{\sum_{j=1}^{|\mathcal{D}_s|} e^{sim(w_i^{X_i}, w_j^{Y_i})/\tau}} \tag{6}$$

其中$|\mathcal{D}_s|$代表的是采样的翻译对数量，τ是一个温度参数。最终，整个多子空间对比微调总损失定义如下：

$$\mathcal{L} = \mathcal{L}_{inter} + \mathcal{L}_{intra} \tag{7}$$

3.4 多子空间动态更新

单轮子空间分配可能会引入来自初始解的噪声，从而使跨语言词嵌入的学习陷入局部最优。因此，我们提出在更新$\mathbf{W}_X$和$\mathbf{W}_Y$的过程中，动态调整目标语言中每个词的子空间分配。

具体来说，源语言$X = \{X_1, X_2, \ldots, X_K\}$中的多个子空间在聚类过程完成后是固定的。对于目标语言中的词w_i^Y，其对应的源语言翻译w_i^X是通过基于对跨语言词嵌入$\mathbf{W}_X X$和$\mathbf{W}_Y Y$检索

得到的。单词w_i^X的子空间索引会分配给w_i^Y。然而，随着$\mathbf{W}_X$和$\mathbf{W}_Y$的更新变化，翻译对也会相应地发生变化。为了保持一致性且避免算法陷入局部最优，每当单词w_i^Y的翻译发生变化时，其对应的子空间归属也随着翻译对的调整而调整。

进一步从计算需求看，DM-BLI 动态更新的时间复杂度为$O(n \cdot \log n)$，其中n 为单词规模。复杂度推导逻辑为：Wasserstein 距离近似计算和CSLS 搜索引入$\log n$ 因子，批处理与子空间优化将计算分担，迭代次数L 进一步调节复杂度，最终呈现线性与对数级叠加的增长趋势。实验数据验证了这一复杂度：当$n = 20000$ 时，动态更新耗时为4946.05 秒；当$n = 40000$ 时，耗时为4930.46 秒；当$n = 75000$ 时，耗时为5124.40 秒。耗时随n 增加的轻微变化符合$O(n \cdot \log n)$的线性增长趋势，反映了GPU 优化和子空间划分对计算效率的影响，也让我们更清晰动态更新在实际执行时的资源消耗特征。由于动态更新的时间复杂度为$\boldsymbol{O(n \cdot \log n)}$，耗时增幅仅约3.6%，为非指数级别增加。这表明方法在处理超大词汇量时是可行的，尽管内存需求和迭代次数可能增加。

正如之前提到的，整个微调过程将逐个子空间进行，顺序按照源语言子空间中单词数量的降序排列。对于目标语言中的每个子空间Y_i，整个动态更新过程将持续至收敛。收敛的判定条件依据Y_i中当前轮次和上一轮次目标词的重叠比例决定。当重叠比例达到α，则认为子空间的动态更新已收敛。此外，一旦一个子空间达到了收敛，该子空间中的词将保持不变。整个方法的流程见算法1。

算法 1 动态多子空间对齐框架

输入: 单词词嵌入$X \in R^{N \times d}$ 和$Y \in R^{M \times d}$
输出: 子空间映射矩阵$\{W_{X_i}\}_{i=1}^K$, $\{W_{Y_i}\}_{i=1}^K$

仅对X 进行重构和聚类，得到K 个子空间$\{X_i\}_{i=1}^K$，其中$X_i \in R^{N_i \times d}$
通过初步对齐，获取初始的映射矩阵$W_{X_{init}}$ 和$W_{Y_{init}}$
根据初始解，完成目标语言Y 的子空间分配
for $i = 1$ **to** K **do**
    初始化映射矩阵$W_{X_i} \leftarrow W_{X_{init}}$, $W_{Y_i} \leftarrow W_{Y_{init}}$
    while 未收敛 **do**
        优化对比学习损失$\mathcal{L}$，更新W_{X_i} 和W_{Y_i}
        保持源语言子空间X_i 的分配不变
        依据更新后的W_{X_i} 和W_{Y_i}，更新Y_i 内的单词分配
    end while
end for
输出$\{W_{X_i}\}_{i=1}^K$ 和$\{W_{Y_i}\}_{i=1}^K$

4 基于DM-BLI的跨语言能力评估应用

本小节提出基于DM-BLI的大语言跨语言能力评估,用于更准确地测量LLMs的跨语言能力。现有基于BLI任务的LLM跨语言能力评估方法，通常将BLI词典中预定义的翻译对视为唯一的“正确答案”，以此来判断模型生成翻译的准确性。然而，此类方法存在显著的局限性，主要体现在无法有效覆盖一词多义现象。例如，英语词汇“bank”既可表示“银行”，亦可指“河岸”，而传统BLI词典往往仅收录最常见的翻译词对，无法涵盖所有语义合理的翻译结果。这种词义覆盖不足导致在评估LLMs时，如果模型给出的翻译并不出现在BLI词典中，即使其语义是合理的，也会被误判为错误。显然无法客观正确评估LLMs的跨语言能力。

因此，本文提出一种结合DM-BLI词典扩展与词向量空间语义特征的评估方法。该方法通过embedding层面的语义相似性度量，对模型生成的翻译进行验证，从而识别出未被词典覆盖但在语义上合理的翻译词。相较于传统基于静态词典匹配的评估方法，本文方法在评估的鲁棒性与语义泛化能力方面具有显著优势，能够更加全面、准确地衡量LLMs在跨语言语义映射任务中的表现。

4.1 评估数据集和提示词

基于上述目标，我们构建了一个多语言—中文映射的评估数据集。具体流程如下：我们首

先基于DM-BLI词典提取了1000余组中文翻译为单字的英文词对（对英文翻译词去重处理）。随后，在其余11种语言中进一步检索与这些英文翻译对应的词汇，最终构建出由12种语言（含英文）映射到中文的跨语言词汇对集合。为保障模型生成行为的可控性，我们设计了两类prompt完成任务诱导：

翻译任务：模型被要求将目标语言单词翻译为中文。此类prompt设计借鉴Few-shot提示策略，通过提供多个上下文样例引导模型生成指定语言翻译。提示格式如图3(a)：

```
Deutsch: "ausserdem" - 中文: "也"
Deutsch: "doch" - 中文: "但"
Deutsch: "einen" - 中文: "一"
Deutsch: "neuer" - 中文: "新"
Deutsch: "ihr" - 中文: "
```

(a) 德语- 中文

```
中文: "也" - 中文: "也"
中文: "但" - 中文: "但"
中文: "一" - 中文: "一"
中文: "新" - 中文: "新"
中文: "你" - 中文: "
```

(b) 中文- 中文

图 3: 翻译、重复任务提示格式示例

重复任务：模型被要求重复给定的中文单词。此类任务验证了LLMs在中文语境中生成能力的输出稳定性，为后续相似度分析提供对照组。提示格式如图3(b)：

4.2 跨语言能力评估

我们基于上述两类任务的模型输出，设计以下评估流程与指标：

1）输出嵌入提取：针对每个翻译或重复任务样本，提取模型生成文本的首个token的嵌入向量，对于重复任务，仅保留生成内容与提示完全一致的样本，以确保嵌入的语义一致性和分析的可比性。**2)语义相似度计算：**采用余弦相似度作为语义一致性度量，计算翻译任务中的生成词与重复任务中生成词之间的余弦相似度及每种语言的平均余弦相似度AvgCosine。**3)输出准确率：**以生成词与参考词完全一致的样本占比，定义为生成准确率Accuracy。**4)相关性分析与拟合建模：** 皮尔逊相关系数（Pearson Correlation Coefficient）用于衡量平均余弦相似度与输出准确率之间的线性相关性；线性回归分析技术构建如下模型以拟合二者之间的关系,并通过R^2值评估拟合程度与预测效果：

$$\text{Accuracy} = a \cdot \text{AvgCosine} + b \tag{8}$$

5 实验设置

我们在12种语言对上评估我们的框架，涵盖了5种资源丰富的语言对：西班牙语（ES）、德语（DE）、俄语（RU）、阿拉伯语（AR）、日语（JA）和中文（ZH），这些语言均与英语（EN）进行跨语言对齐，以及5种低资源语言对：芬兰语（FI）、印地语（HI）、土耳其语（TR）、印尼语（ID）、保加利亚语（BG）和加泰罗尼亚语（CA），同样也与英语（EN）进行跨语言对齐。根据先前的研究(Mohiuddin et al., 2020b; Tian et al., 2022)，我们在实验中也选择相对低资源的语言对作为低资源设置。

5.1 数据集

我们使用在维基百科上训练的fastText向量(Bojanowski et al., 2016)作为单语词嵌入，以及Facebook发布的MUSE双语词典(Conneau et al., 2018)作为基准词典。MUSE提供了110个双语词典，每个词典包含每种语言中使用频率最高的6,500个单词，分为包含1,500个单词的测试集和包含5,000个单词的训练集。

5.2 基线模型

本文的基线模型分为有监督和无监督两组设置，我们在实验中运行了每个基线模型的公开代码。

有监督BLI: （1）MUSE(Conneau et al., 2018)：通过最小化监督翻译对之间的欧氏距离来学习正交映射。（2）VecMap(Artetxe et al., 2018)：一个学习双语词嵌入的通用多步框架，包括白化、正交映射、重新加权、去白化和降维。（3）BLISS(Patra et al., 2019)：是一种半监

督方法，其中损失函数包含一个弱正交约束和反向翻译。（4）CL-BLI(Li et al., 2022)：一个强大且有效的两阶段对比学习框架，结合静态和上下文词嵌入。

无监督BLI: （1）MUSE(Conneau et al., 2018)：无监督MUSE基于普氏分析(Procrustes analysis，PA)执行对齐算法归纳种子词典，使用对抗的方式学习映射矩阵。（2）VecMap(Artetxe et al., 2018)：无监督VecMap使用单语内词相似性信息来诱导初始解决方案。（3）Ad.(Mohiuddin and Joty, 2019)：是一种对抗自编码器框架，先通过自编码器将词嵌入映射至潜在嵌入空间中，后进行对抗学习训练。（4）BLOOM-7B：这是一个仅解码器的Transformer语言模型，支持46种自然语言。我们在实验中使用了7B参数版本。（5）LLaMA-13B(Touvron et al., 2023)：这是一个仅有解码器结构的Transformer语言模型，支持20种语言。我们在实验中使用了13B参数版本。（6）GPT-3.5(Brown et al., 2020)：这是一个仅有解码器结构的Transformer语言模型，拥有175B参数，支持38种语言。我们在实验中使用了GPT-3.5-turbo版本。

5.3 评估指标与参数设置

我们选取每种语言中最频繁的75,000个词的fastText词嵌入作为输入，并进行标准化处理，包括：长度归一化、均值中心化和再次归一化。

簇间对比学习中，仅选取置信度高于0.45的词表示子空间分布；簇内对比学习中，从按置信度排序的翻译对中采样，采样比例$Sample_\alpha$设为20%。收敛判断标准为：若某子空间本轮与上一轮的词汇重叠率达到95%（即$\alpha = 0.95$），则认为已收敛。

在使用大语言模型进行BLI任务时，我们采用Li等人(Li et al., 2023)提出的最佳提示模板。LLaMA-13B的提示格式为："Translate from L_x to L_y: $w^x \Rightarrow$"；GPT-3.5使用："Translate the L_x word into L_y: "。

BLI任务的评估指标为Precision@1，即正确翻译是否出现在最近邻目标词集合中。候选集合通过CSLS算法从跨语言词向量中获取。我们使用MUSE双语词典(Mohiuddin et al., 2020b)作为参考，将每个模型生成的双语词典与之比对，评估其在BLI任务中的表现。

6 实验结果及讨论

6.1 高资源和低资源语言对的BLI实验结果

表 1: 不同模型在高低资源语言对上的双语词典归纳准确率对比

模型	高资源语言							低资源语言						
	Es-En	De-En	Ru-En	Ar-En	Ja-En	Zn-En	Avg.	Fi-En	Hi-En	Tr-En	Id-En	Bg-En	Ca-En	Avg.
有监督														
MUSE	67.80	63.14	53.23	44.33	0.14	8.29	39.50	46.50	25.65	39.80	35.50	39.28	46.19	38.80
BLISS	68.46	63.49	54.88	45.70	0.01	6.43	39.80	49.90	28.17	41.40	38.40	42.21	47.26	41.30
VecMap	71.70	66.46	59.58	51.54	37.14	42.50	54.80	58.12	34.07	49.37	44.72	49.13	54.35	48.30
CL-BLI	**73.02**	**69.00**	61.31	**53.14**	35.07	42.44	55.70	57.78	32.62	48.52	43.43	47.34	53.89	47.30
DM-BLI	72.87	68.28	**61.61**	52.33	**41.03**	**44.83**	**56.80**	**60.29**	**35.57**	**53.09**	**48.24**	**50.80**	**56.47**	**50.70**
无监督														
MUSE	67.89	63.27	50.49	0.03	0.09	0.01	30.30	0.05	0.00	36.82	36.35	38.31	46.07	26.30
VecMap	72.00	67.17	56.42	47.43	26.62	33.39	50.50	54.71	28.19	48.92	45.65	45.69	53.52	46.10
Ad.	71.93	66.63	55.50	0.00	0.00	0.00	32.30	0.45	0.01	46.69	0.09	0.03	53.06	16.70
BLOOM-7B	52.50	38.34	26.06	32.67	21.34	34.35	34.20	23.43	28.30	30.82	45.45	16.75	43.89	31.40
LLaMA-13B	60.58	57.80	64.44	22.13	38.56	32.28	46.00	40.98	30.68	44.90	48.63	56.86	48.83	45.10
GPT-3.5	68.17	63.07	**74.15**	**65.94**	**71.80**	**65.12**	**68.00**	**60.37**	**56.11**	**54.49**	48.37	**67.51**	45.15	**55.3**
DM-BLI	**72.94**	**68.67**	58.91	48.58	32.42	37.34	53.10	57.48	30.80	51.98	**48.81**	47.63	**56.15**	48.80

表1总结了DM-BLI在高资源和低资源语言对上的监督及无监督双语词典归一化任务的结果。在监督任务中，DM-BLI在高资源语言对上平均性能达56.8%，比最优基线CL-BLI 55.7%提升1.1%，在所有语言对上均表现最佳或次优，特别在Ja-En上提升5.96%，从35.07%提升至41.03%。在低资源语言对上，DM-BLI平均性能为50.7%，优于VecMap 48.3%，展现稳健性。在无监督任务中，DM-BLI在高资源语言对上平均性能为53.1%，优于VecMap 50.5%，但低于GPT-3.5 68.0%。在低资源语言对上，DM-BLI平均性能为38.8%，在四种语言对上超越基线，但不及GPT-3.5 55.3%，仍具一定鲁棒性。GPT-3.5在平均性能上表现更优，其原因一方

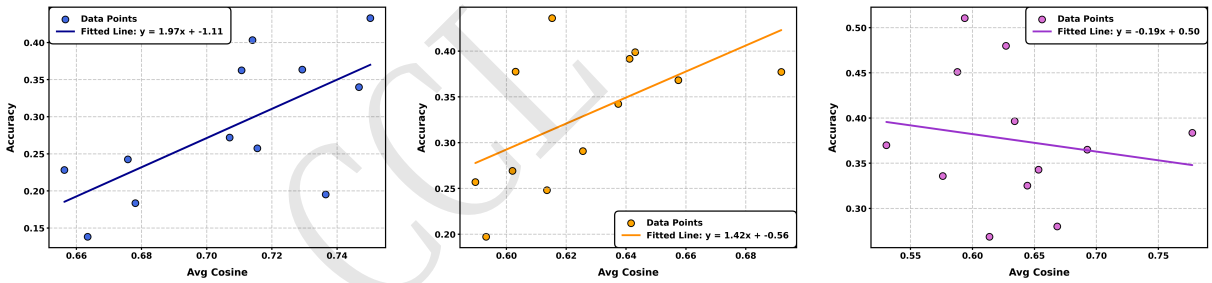
面是因为GPT-3.5拥有1750亿参数，基于海量互联网文本预训练，覆盖多种语言和领域，特别是在无监督和少样本数据上泛化能力强；另一方面，我们通过检查初始单语词嵌入质量发现，DM-BLI 在许多语言对上性能欠佳，例如Zh-En，是因为中文与英语语义空间异构性高，初始词嵌入语义表示不足。DM-BLI 表现优于GPT-3.5的语言对有Ca-En、Es-En、De-En等，原因是这些语言对与英语共享较多语义特征，说明DM-BLI 的子空间优化和Wasserstein 距离方法能有效利用这一特性提升对齐精度，而GPT-3.5 的泛化优势在此场景下未充分发挥。相比之下，BLOOM-7B和LLaMA-13B表现不佳，落后于传统方法。

总体而言，DM-BLI在监督任务中不仅在高资源语言对中展现出显著的跨语言适应能力，也在低资源语言环境下保持了较高的稳定性和泛化能力，优于现有的传统方法如CL-BLI和VecMap。在无监督任务中，DM-BLI整体优于传统映射方法，特别是在高资源语言对中表现出色，体现出较强的内在表示学习能力。然而，在低资源语言对的无监督设定下，尽管DM-BLI超越了多个基线模型，但与GPT-3.5等大规模预训练模型仍存在显著差距。未来工作可进一步结合大模型的上下文建模能力与DM-BLI的结构优势，以提升其在低资源跨语言任务中的表现。

6.2 基于DM-BLI的跨语言能力评估应用实验

表 2: 不同模型在各语言对下的指标对比

语言对	LLaMA		Mistral		Qwen	
	Accuracy	AvgCosine	Accuracy	AvgCosine	Accuracy	AvgCosine
Ar-Zh	0.18	**0.68**	0.25	0.61	**0.34**	0.65
Bg-Zh	0.23	**0.66**	0.26	0.59	**0.33**	0.64
Ca-Zh	0.24	**0.68**	**0.38**	0.60	0.37	0.53
De-Zh	0.34	**0.75**	**0.37**	0.66	**0.37**	0.69
En-Zh	0.36	0.71	**0.38**	0.69	**0.38**	0.78
Es-Zh	0.36	**0.73**	0.39	0.64	**0.45**	0.59
Fi-Zh	**0.27**	**0.71**	**0.27**	0.60	**0.27**	0.61
Hi-Zh	0.14	**0.66**	0.20	0.59	**0.34**	0.58
Id-Zh	0.26	**0.72**	0.34	0.64	**0.40**	0.63
Ja-Zh	0.43	**0.75**	0.40	0.64	**0.51**	0.59
Ru-Zh	0.40	**0.71**	0.44	0.62	**0.48**	0.63
Tr-Zh	0.20	**0.74**	**0.29**	0.63	0.28	0.67



(a) LLaMA-2-7B (b) Mistral-7B-Instruct-v0.3 (c) Qwen2.5-7B-Instruct

图 4: 线性回归拟合Accuracy与AvgCosine之间的关系：(a) LLaMA-2-7B（$R^2 = 0.4483$）；(b) Mistral-7B-Instruct-v0.3（$R^2 = 0.3235$）；(c) Qwen2.5-7B-Instruct（$R^2 = 0.0268$）。

在多语言语义一致性对比实验中，本文引入的重复-翻译任务嵌入对齐方法展现出良好的跨模型适用性与显著的分析能力。如表2所示，LLaMA-2-7B、Mistral-7B-Instruct-v0.3和Qwen2.5-7B-Instruct在各语言对上的翻译准确率（Accuracy）和平均余弦相似度（AvgCosine）呈现出不同的性能表现。例如，LLaMA在Es-Zh语言对中，Accuracy为0.36，AvgCosine为0.73，表现出较高的一致性；而Qwen在En-Zh语言对中，AvgCosine达0.78但Accuracy仅0.38，反映了模型间在语义对齐上的差异。

为进一步分析Accuracy与AvgCosine之间的相关性，本文采用皮尔逊相关系数进行计算，并通过线性回归拟合进行可视化展示，如图4。LLaMA-2-7B的皮尔逊相关系数$r = 0.6695$，p值约为0.017，表明其Accuracy与AvgCosine之间存在显著的正向线性相关，线性回归拟合方程为$y = 1.97x - 1.11$，$R^2 = 0.4483$，如图4(a)所示。Mistral-7B-Instruct-v0.3的皮尔逊相关系

数$r = 0.5687$，p值约为0.053，接近显著性阈值（$p < 0.05$），回归拟合方程为$y = 1.42x - 0.56$，$R^2 = 0.3235$，显示中等程度的正向相关，如图4(b)所示。Qwen2.5-7B-Instruct的皮尔逊相关系数$r = -0.1638$，p值约为0.61，表明无显著相关性，其回归拟合方程为$y = -0.19x + 0.50$，$R^2 = 0.0268$，如图4(c)所示。

通过对三个模型的综合分析可知，LLaMA在多语言表示对齐方面展现出最强的语义一致性，其重复任务嵌入结构在捕捉跨语言语义关系上具有较高的稳定性与可预测性。Mistral模型次之，在特定语言对，如Ca-Zh、En-Zh中，依然具备较强的对齐能力，适合用于语义相似性驱动的翻译任务评估。相较之下，Qwen模型在语义空间构建策略上表现出明显差异，可能更依赖上下文驱动机制或采用非对称语言建模结构，导致重复生成与翻译输出之间缺乏稳定的一致性。实验结果表明，重复任务嵌入与翻译输出之间的一致性指标可作为衡量多语言模型语义对齐能力的有效手段，并进一步揭示了不同模型架构在语义建模机制上的差异性。

6.3 消融实验

我们设计了消融实验以证明对比学习模块的重要性，结果如表3 所示。移除对比学习后，准确率从0.791降至0.766（下降约2.47%），匹配词数从1186减至1149；子空间分配由均匀分布（如453、388）变为集中分布（如42、28），总词汇覆盖从70227增至74662。这表明对比学习通过正负样本优化显著增强了子空间内词义区分度，其移除导致对齐精度降低，因缺乏对比损失引导，模型收敛至更小词汇子集，从而未分配词汇减少但覆盖率提升。上述结果有力验证了对比学习在提高准确率（尤其最近邻匹配）中的关键作用。

表 3: 有无对比学习模块的消融实验结果对比表

实验条件	使用对比学习模块	移除对比学习模块
准确率(Acc)	0.791	0.766
匹配词数(k=5)	1186	1149
子空间分配(示例)	453个，388个	42个，28个
总词汇覆盖	70227	74662

6.4 子空间对齐效果差异

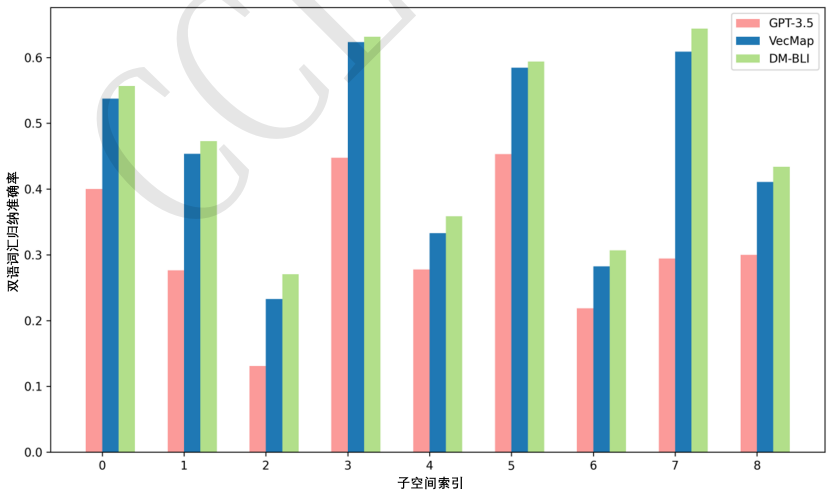


图 5: 在不同的英语子空间中，Ca-En的无监督BLI(Percision@1)

如图5可见，DM-BLI 在Ca-En 各子空间的Precision@1 虽整体优于VecMap，但子空间性能差异显著（如子空间3、7 准确率远高于其他）。这是因为Ca-En 作为低资源语言对，数据样本本身稀缺，部分子空间因样本量不足、语义异质性强，无法支撑有效对齐，反映低资源场景下子空间对齐的脆弱性。

6.5 子空间数量鲁棒性验证

由于无参数聚类法对数据的分布和噪声较为敏感，我们对从VecMap 和DM-BLI 派生的采样CLWE 进行t-SNE 可视化，如图6所示。结果显示，子空间数量在达到一定值后差异不大，趋于稳定。图中不同子空间数量（5、7、9）的嵌入分布聚类结构变化不大，验证了该方法对子空间数量的鲁棒性，表明该方法对噪声和分布的敏感性影响有限。FINCH 无需预设聚类数K，通过层次化方法自适应确定集群数量，适用于复杂数据分布;其基于连接性优化的聚类更适合非球形或不均匀密度数据。

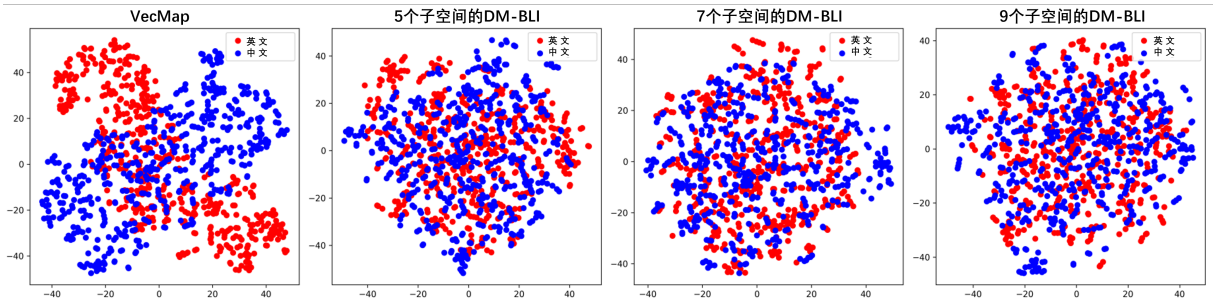


图 6: 从VecMap 和DM-BLI 派生的采样CLWE 进行t-SNE 可视化，其中从DM-BLI 派生的CLWE 可视化基于不同数量的多个子空间

7 总结与展望

本文提出了一种用于无监督双语词典归纳的动态多子空间对齐框架，称为DM-BLI。我们的方法摒弃单一的全局映射，提出将词嵌入空间分割成多个子空间，在考虑全局信息的基础上在多个子空间上实现对齐，从而实现更细粒度且更准确的跨语言词嵌入对齐。实验表明，与基线方法相比，我们的方法可以显著提高双语词汇归纳的性能，尤其是在远距离和低资源语言对上。同时，本文进一步基于DM-BLI构建跨语言评估应用，以提升大语言模型跨语言能力评估的语义覆盖与鲁棒性。我们提出结合翻译任务与重复任务的评估框架，通过对模型输出嵌入进行语义相似度度量，识别出传统词典无法覆盖但语义合理的翻译词，解决了现有方法在一词多义、多语义表达等情境下的评估偏差问题。多模型对比实验表明，该方法在多个语言对上均可有效刻画模型的语义一致性表现，准确率与平均余弦相似度之间的相关性进一步揭示了不同模型在语义建模策略上的差异。本方法为LLMs跨语言语义能力评估提供了一种更具解释力和区分性的分析路径，也为后续模型设计与调优提供了参考依据。

致谢

本成果受国家社科基金项目“语言安全视域下大语言模型的价值观生成机制和风险应对研究”（No.24CYY107），中央高校基本科研业务费专项“生成式人工智能的价值观对齐研究”（No.2024TD001）基金资助。

参考文献

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

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*, page 1881–1890. Brussels, Belgium.

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.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 9432–9452. Toronto, Canada.

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 and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, page 7059–7069. Red Hook, NY, USA.

A Conneau, G Lample, M Ranzato, et al. 2018. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations*, pages 1 – 14. Vancouver, Canada.

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)*, page 4171–4186. Minneapolis, Minnesota.

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)*, page 227–23. Online.

M Faruqui and C Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462 – 471. Gothenburg, Sweden.

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*, page 7548–7555. Online.

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

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 – 46. Online.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 25th international conference on Machine learning(volume 15)*, page 748–756.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, PengJen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Yuehui Han, Le Hui, Haobo Jiang, Jianjun Qian, and Jin Xie. 2022. Generative subgraph contrast for self-supervised graph representation learning. In *Proceedings of the 17th European Conference on Computer Vision*, page 91–107. Berlin, Heidelberg.

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.

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*, page 9577–9599. Singapore.

第二十四届中国计算语言学大会论文集，第792页-第806页，济南，中国，2025年8月11日至14日。

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*, page 1442–1459. Online and Punta Cana, Dominican Republic.

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*, page 281–297. Oakland, CA, USA.

Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, Carey Priebe, and Philipp Koehn. 2022a. 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. AbuDhabi, United Arab Emirates.

Kelly Marchisio, Neha Verma, Kevin Duh, and Philipp Koehn. 2022b. Isovec: Controlling the relative isomorphism of word embedding spaces. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 6019–6033. Abu Dhabi, United Arab Emirates.

Leah Michel, Viktor Hangya, and Alexander Fraser. 2020. Exploring bilingual word embeddings for hiligaynon, a low-resource language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, page 2573–2580. Online.

T Mikolov, K Chen, G Corrado, et al. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*, pages 1 – 12. Scottsdale, Arizona.

Tasnim Mohiuddin and Shafiq Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. 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)*, page 3857–3867. Minneapolis, Minnesota.

Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. 2020a. 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)*, page 2712–2723. Online.

Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. 2020b. 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)*, page 2712–2723. Online.

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.

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*, page 184–193. Florence, Italy.

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*, page 3476–3485. Online.

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*, page 8934–8943. Long Beach, CA, USA.

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)*, page 778–788. Melbourne, Australia.

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

Zhoujin Tian, Chaozhuo Li, Shuo Ren, Zhiqiang Zuo, Zengxuan Wen, Xinyue Hu, Xiao Han, Haizhen Huang, Denvy Deng, Qi Zhang, and Xing Xie. 2022. Rapo: An adaptive ranking paradigm for bilingual lexicon induction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 8870–8883. Abu Dhabi, United Arab Emirates.

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.

I Vuli and M F Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bi-lingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 719 – 725. Beijing, China.

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)*, page 7222–7240. Online.

Haozhou Wang, James Henderson, and Paola Merlo. 2021. Multi-adversarial learning for cross-lingual word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 463–472. Online.

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)*, page 1959–1970. Vancouver, Canada.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, page 41092–41110. Hawaii, USA.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023b. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, page 5597–5607. New York, NY, USA.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, et al. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, page 2765–2781. Mexico City, Mexico.

W Y Zou, R Socher, D Cer, et al. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393 – 1398. Seattle, Washington, USA.