·人工智能与模式识别· 文章编号：1000-3428(2023)01-0082-10 文献标志码：A 中图分类号：TP391

# 融合词簇约束的汉越跨语言词嵌入

武照渊[1,2]，余正涛[1,2]，黄于欣[1,2]

（1.昆明理工大学 信息工程与自动化学院，昆明 650500；2.云南省人工智能重点实验室，昆明 650500）

**摘 要：**针对传统跨语言词嵌入方法在汉越等差异较大的低资源语言上对齐效果不佳的问题，提出一种融合词簇对齐约束的汉越跨语言词嵌入方法。通过独立的单语语料训练获取汉越单语词嵌入，使用近义词、同类词和同主题词3种不同类型的关联关系，充分挖掘双语词典中的词簇对齐信息以融入到映射矩阵的训练过程中，使映射矩阵进一步学习到不同语言相近词间具有的一些共性特征及映射关系，根据跨语言映射将两种语言的单语词嵌入映射至同一共享空间中对齐，令具有相同含义的汉语与越南语词嵌入在空间中彼此接近，并利用余弦相似度为空间中每一个未经标注的汉语单词查找对应的越南语翻译构建汉越对齐词对，实现跨语言词嵌入。实验结果表明，与传统有监督及无监督的跨语言词嵌入方法Multi_w2v、Orthogonal、VecMap、Muse相比，该方法能有效提升映射矩阵在非标注词上的泛化性，改善汉越低资源场景下模型对齐效果较差的问题，其在汉越双语词典归纳任务P@1和P@5上的对齐准确率相比最好基线模型提升了2.2个百分点。

**关键词：**汉越双语；低资源语言；跨语言词嵌入；词簇对齐；多粒度约束

**开放科学（资源服务）标志码（OSID）：**

**中文引用格式：**武照渊，余正涛，黄于欣.融合词簇约束的汉越跨语言词嵌入[J].计算机工程，2023，49(1)：82-91.

**英文引用格式：**WU Z Y，YU Z T，HUANG Y X.Chinese-Vietnamese cross-lingual word-embedding combined with word cluster constraints[J].Computer Engineering，2023，49(1)：82-91.

# Chinese-Vietnamese Cross-Lingual Word-Embedding Combined with Word Cluster Constraints

WU Zhaoyuan[1,2]，YU Zhengtao[1,2]，HUANG Yuxin[1,2]

（1.Faculty of Information Engineering and Automation，Kunming University of Science and Technology，Kunming 650500，China；2.Yunnan Key Laboratory of Artificial Intelligence，Kunming 650500，China）

**【Abstract】**To solve for the poor alignment effect of the traditional cross-lingual word-embedding method in low-resource languages such as Chinese-Vietnamese，this paper proposes a Chinese-Vietnamese cross-lingual word embedding method with word cluster alignment constraints.First，Chinese and Vietnamese monolingual word embeddings are obtained via training on an independent monolingual corpus.Subsequently，three different types of association relationships including synonyms，similar words，and same subject words are used to completely mine the word cluster alignment information in the bilingual dictionary and integrate it into the training process of the mapping matrix.This allows the mapping matrix to further learn some common features and mapping relationships between similar words in different languages.Second，the monolingual word embeddings of the two languages are mapped onto a shared space through cross-lingual mapping to ensure that the Chinese and Vietnamese word embeddings with the same meaning are close to each other in the space.Finally，the cosine similarity is used to find the corresponding Vietnamese translation for each non-labeled Chinese word in the space，and Chinese-Vietnamese aligned word pairs are constructed to realize cross-lingual word embedding.The experimental results show that the proposed method is different from traditional supervised and unsupervised cross-lingual word-embedding methods such as Multi_w2v，Orthogonal，VecMap，and Muse，and can effectively improve the generalization of the mapping matrix with non-labeled words and poor effect of model alignment in low-resource languages such as Chinese-Vietnamese.Moreover，its alignment accuracy in the Chinese-Vietnamese bilingual dictionary induction tasks P@1 and P@5 is improved by 2.2 percentage points compared with that of the best baseline model.

**基金项目：**国家自然科学基金（61732005，U21B2027，61972186，61866020，61866019）；云南省重大科技专项（202002AD080001，202103AA080015）；云南省高新技术产业专项（201606）。

**作者简介：**武照渊（1997—），男，硕士研究生，主研方向为自然语言处理、智能信息处理；余正涛（通信作者），教授、博士、博士生导师；黄于欣，副教授、博士。

**收稿日期：**2021-11-30 **修回日期：**2022-01-27 **E-mail：**ztyu@hotmail.com

【Key words】Chinese-Vietnamese bilingual; low-resource language; cross-lingual word embedding; word cluster alignment; multi-granularity constraints
**DOI**:10. 19678/j. issn. 1000-3428. 0063407

## 0　概述

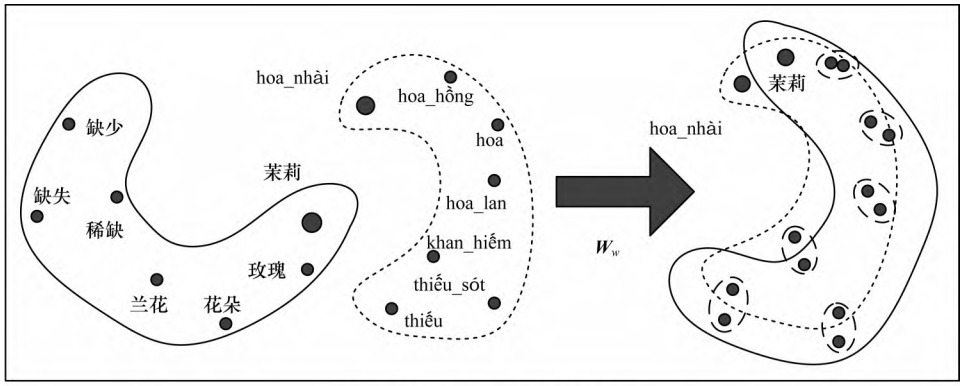
跨语言词嵌入将不同语言具有相同含义的词映射至同一空间中对齐,是跨语言文本分类[1-3]、跨语言情感分析[4-5]、机器翻译[6-8]、跨语言实体链接[9-10]等任务的基础,具有重要的应用价值。

汉越跨语言词嵌入是面向低资源语言的双语词嵌入任务,目前低资源跨语言词嵌入方法主要包括无监督、半监督和有监督3类。无监督方法利用不同语言单语嵌入空间的相似性,无需标注数据即可学习映射矩阵实现对齐。文献[11]通过初始化一个映射矩阵作为生成器,使映射后的源语言词嵌入更加接近目标语言词嵌入以欺骗鉴别器,利用对抗思想迭代优化映射矩阵。文献[12]将两种语言的词嵌入空间视作两个分布,通过最小化分布间的沃瑟斯坦距离实现跨语言词嵌入。文献[13]则是将沃瑟斯坦距离同对抗训练相结合,使生成器构建出的源语言词嵌入更加接近目标语言词嵌入,以提升对抗方法在低频词上的性能。半监督方法相比无监督方法引入了少量词对齐信息用于提升映射的准确性,主要包括两种策略:一种是使用小规模平行词对作为监督信号来解决映射矩阵在迭代训练过程中初始化阶段的不足;另一种则是将无监督模型学习到的嵌入空间相似性同有监督模型相结合来提升对齐效果。文献[14]使用少量的双语词对学习初始映射矩阵,并将映射得到的翻译词对作为扩展数据迭代学习新的映射矩阵。文献[15]发现统计翻译模型仅凭少量平行语料即可实现不同语言高频词间的准确对齐,而映射模型则能通过大规模的单语语料实现低频词间的对齐,因此提出一种融合统计与映射方法的跨语言词嵌入模型。文献[16]发现无监督模型更倾向于学习大范围的空间对齐,而有监督模型更善于学习词对间的精确对齐,基于这一思想提出将无监督损失同有监督损失进行联合优化来提升映射效果。目前已有的半监督与无监督方法在相近语言上取得了不错的效果,例如英语-西班牙语由于词根、构词方式上具有相似性,两者词嵌入空间包含大量的共现词与同构词(如“possible”与“posible”),仅凭少量标注数据就能实现较好的对齐。然而,词源学上差异较大的语言通常单语嵌入空间之间的相似性也较低[17],对于汉越这种差异较大的语言,无监督和半监督的方法对齐效果不佳[18]。当前,通过双语词典学习映射矩阵的有监督方法可以有效提升远距离语言上的对齐效果,如文献[19]提出使用双语词典作为监督信号,通过最小化词典词对间的欧氏距离平方和来学习映射矩阵,在英语-捷克语上取得了不错的效果。一些后续研究则在此基础上通过引入归一化处理[20]和为映射矩阵添加正交约束[21]来进一步提升映射的准确性。为缓解远距离语言间语法差异带来的影响,文献[22]提出分别为两种语言学习单独的映射矩阵,将不同语言词嵌入映射至一个同语言无关的共享空间中来最大化其相似度。文献[23]通过引入语言学中语言家族树的概念,利用层次化映射将文献[22]中的方法扩展至多语言任务上。然而,传统有监督方法仅使用词典中的词对齐信息学习映射矩阵,汉越作为低资源语言对其双语词典在规模及质量上同资源富集型语言(如:汉语-英语)仍有较大差距,导致学习到的映射矩阵在词典外的非标注词上对齐效果不佳。

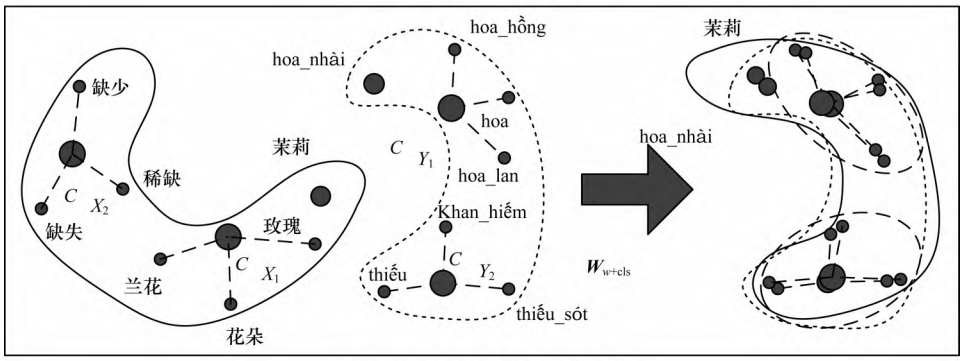
本文提出一种融合词簇对齐约束的汉越跨语言词嵌入模型,用于改善低资源场景下汉越双语空间的对齐效果。使用不同类型的关联关系充分挖掘双语词典中蕴含的词簇对齐信息,通过构建词与词簇两种粒度的联合损失将其融入到映射矩阵的训练中,以提升映射矩阵在非标注词上的泛化性,在此基础上结合汉越双语词典的特点设置了近义词、同类词和同主题词3种类型的对齐词簇,并通过实验研究不同类型词簇及其组合对模型效果的提升。

## 1　汉越跨语言词嵌入模型

基于双语词典学习跨语言词嵌入的关键步骤是学习单语嵌入空间之间的映射关系[24],传统有监督模型通常使用双语词典中的词对齐信息学习映射矩阵,对齐两种语言的嵌入空间,如图1(a)所示。但汉越作为低资源语言对缺乏大规模的双语词典,导致学习到的映射矩阵 $\boldsymbol{W}_w$ 在双语词典外的非标注词上泛化性较弱,无法准确对齐双语空间。例如图中的非标注词“茉莉”,经过映射后仍与对应的越南语翻译“hoa_nhài”距离较远,对齐效果欠佳。实际上,词典中存在一些近义词与同类词,如图1(b)所示,“缺少”、“稀缺”、“缺失”和“兰花”、“花朵”、“玫瑰”这种具有相近含义的词在汉语词嵌入空间中的距离比较接近,可以构建为词簇,且词典中相应的翻译“thiếu”、“khan_hiếm”、“thiếu_sót”和“hoa_lan”、“hoa”、“hoa_hồng”在越南语词嵌入空间中也具有邻近的分布。




(a)仅使用词对齐信息学习映射矩阵




(b)融合词簇对齐信息学习映射矩阵

**图1 融合词簇约束前后的汉越词嵌入空间对齐效果**

**Fig.1 Alignment effect of Chinese and Vietnamese word embedding space before and after with word cluster constraints**

文献[25]提出来自不同语言具有相近含义的词簇在映射后的距离也应接近。因此,本文提出一种融合词簇对齐约束的汉越跨语言词嵌入方法,通过使用词簇对齐数据让映射矩阵 $\boldsymbol{W}_{w+\mathrm{cls}}$ 学习到词簇粒度的映射关系。例如词簇对齐信息“兰花,花朵,玫瑰 | hoa_lan,hoa,hoa_hồng”可以使映射矩阵学习到具有“花朵”含义的汉越单语词嵌入之间的一些共性特征及映射关系(如“花”与“hoa”),使模型在映射过程中尽可能精准地识别和保留这些特征,并通过映射拉近汉越相近语义词在共享空间中的距离。这种共性特征间的映射关系还可以迁移到其他未标注词簇上,通过局部词簇空间的准确对齐提升模型在非标注词上的泛化性,进一步改善低资源环境下汉越整体嵌入空间的对齐效果。例如图1(b)中非标注词“茉莉”与“hoa_nhài”经过映射矩阵 $\boldsymbol{W}_{w+\mathrm{cls}}$ 后,更接近具有“花朵”含义的词簇,词簇空间准确对齐的同时也使两词间的距离更近,更容易实现对齐。

### 1.1 汉越单语词嵌入训练

实现汉越跨语言词嵌入的首要步骤是获取汉越单语词嵌入。模型使用汉语与越南语的单语训练语料作为输入,通过Word2Vec模型[26]训练获取汉语与越南语的单语词嵌入空间 $\boldsymbol{X}\in\mathbb{R}^{n\times d}$,$\boldsymbol{Y}\in\mathbb{R}^{m\times d}$,其中,$n$ 为训练得到的汉语词嵌入个数,$m$ 为越南语词嵌入个数,$d$ 代表词嵌入维度。然后通过汉越双语词典,分别得到词典对应的汉语与越南语词嵌入矩阵 $\boldsymbol{X}'$,$\boldsymbol{Y}'\in\mathbb{R}^{v\times d}$,其中,$v$ 为词典大小,$\boldsymbol{X}_i'$ 与 $\boldsymbol{Y}_i'$ 分别代表双语词典第 $i$ 条词对所对应的汉、越单语词嵌入。

### 1.2 融合词簇对齐约束的映射矩阵训练

汉越双语词典中存在许多近义词、同类词和同主题词可以构建为词簇,近义词是指具有相近含义的词,例如“缺少”、“稀缺”、“缺失”;同类词代表具有某种联系的词,例如“兰花”、“花朵”、“玫瑰”;同主题词指围绕某一主题概念的词,例如表示数字的“零”、“一”、“二”,表示颜色为“红色”、“蓝色”等。模型基于3种不同类型的关联关系充分挖掘双语词典中的词簇对齐信息,数据构建详见2.3节。为了更好地将词簇对齐信息融入映射矩阵的训练过程中,模型通过平均操作将词簇对齐进一步转换为簇心对齐。设现有词簇对齐信息为 $\{X_1',X_2',\cdots,X_s'|Y_1',Y_2',\cdots,Y_t'\}$,$s$ 代表汉语词簇中的词嵌入个数,$t$ 代表越南语词簇中的词嵌入个数,所对应的簇心嵌入 $\boldsymbol{C}_{X_i}$ 与 $\boldsymbol{C}_{Y_i}$ 的构建如式(1)、式(2)所示:

$$\boldsymbol{C}_{X_i}=\frac{1}{s}\sum_{i=1}^{s}\boldsymbol{X}_i' \tag{1}$$

$$\boldsymbol{C}_{Y_i}=\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{Y}_i' \tag{2}$$

这种簇心嵌入与词典中的词嵌入维度相同，因此词对齐数据同词簇对齐数据可以直接融合为训练数据$D_3$，应用到映射矩阵的训练中。$\boldsymbol{C}_X, \boldsymbol{C}_Y \in \mathbb{R}^{l\times d}$分别代表词簇对齐数据中汉语与越南语的簇心嵌入矩阵，$l$代表词簇对齐数据的规模。$\hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}} \in \mathbb{R}^{(v+l)\times d}$分别代表训练数据$D_3$中融合词对齐与词簇对齐数据后的汉语与越南语嵌入矩阵，其中，$\hat{\boldsymbol{X}}=\boldsymbol{X}'+\boldsymbol{C}_X$，$\hat{\boldsymbol{Y}}=\boldsymbol{Y}'+\boldsymbol{C}_Y$。

受语法、构词上的差异及单语训练语料主题不一致等因素的影响，汉语与越南语的词嵌入空间并不同构，而传统跨语言词嵌入方法通常采用从源语言到目标语言的单向映射实现双语空间对齐，并没有考虑这种差异性带来的影响，导致最终获取到的汉越跨语言词嵌入效果不佳。因此，在映射矩阵的训练步骤中，模型的目标为汉语与越南语分别学习两个单独的映射矩阵$\boldsymbol{W}_X$与$\boldsymbol{W}_Y$，从而将两种语言的单语词嵌入映射至一个同语言无关的共享嵌入空间中，以减小汉越语言差异性对模型效果的影响。同时，为两个映射矩阵添加正交约束$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}$，以保证映射后的单语词嵌入性能不变。依据训练数据$D_3$中的对齐关系，词典中对应的汉语与越南语词嵌入经过映射后应尽可能相似。同理，词簇对齐数据中的汉语与越南语簇心嵌入经过映射后也应尽可能相似。这一问题可以通过最小化它们之间欧氏距离的平方和求解，模型在词对齐与词簇对齐上的联合损失函数如式(3)所示：

$$\underset{\boldsymbol{W}_X,\boldsymbol{W}_Y}{\operatorname{argmin}} \sum_i \left\| \boldsymbol{X}_i\boldsymbol{W}_X-\boldsymbol{Y}_i\boldsymbol{W}_Y \right\|^2+\sum_j \left\| \boldsymbol{C}_{X_j}\boldsymbol{W}_X-\boldsymbol{C}_{Y_j}\boldsymbol{W}_Y \right\|^2 \tag{3}$$

不同于仅使用词对齐约束构建映射损失的传统有监督方法，融合词簇对齐约束后的联合损失函数可以使模型进一步学习到汉越相近词间的共性特征及其映射关系，使不同语言具有相近含义的词嵌入在映射后的距离更近。这种距离缩减可以使后期双语词典归纳任务中检索到的越南语候选词同汉语检索词的语义相关性更强，从而提升模型在非标注词上的泛化能力，以弥补低资源环境下词粒度对齐关系学习不充分的问题。因为两个损失中的映射矩阵是共享的，所以通过训练数据$D_3$可以将损失函数进一步简化为：

$$\underset{\boldsymbol{W}_X,\boldsymbol{W}_Y}{\operatorname{argmin}} \sum_i \left\| \hat{\boldsymbol{X}}_i\boldsymbol{W}_X-\hat{\boldsymbol{Y}}_i\boldsymbol{W}_Y \right\|^2 \tag{4}$$

当$\boldsymbol{W}_X$与$\boldsymbol{W}_Y$满足正交约束时，最小化欧氏距离的平方和可以等价于最大化点积，如式(5)所示：

$$\underset{\boldsymbol{W}_X,\boldsymbol{W}_Y}{\operatorname{argmax}} \operatorname{Tr}(\hat{\boldsymbol{X}}\boldsymbol{W}_X\boldsymbol{W}_Y^{\mathrm{T}}\hat{\boldsymbol{Y}}^{\mathrm{T}}) \tag{5}$$

其中：Tr(*)为迹运算，代表矩阵主对角线上所有元素之和，该问题的最优正交解为$\boldsymbol{W}_X=\boldsymbol{U}$，$\boldsymbol{W}_Y=\boldsymbol{V}$，其中$\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}=\hat{\boldsymbol{X}}^{\mathrm{T}}\hat{\boldsymbol{Y}}$为$\hat{\boldsymbol{X}}^{\mathrm{T}}\hat{\boldsymbol{Y}}$的SVD解。

### 1.3　跨语言映射

基于1.2节中获得的汉语与越南语的映射矩阵$\boldsymbol{W}_X$与$\boldsymbol{W}_Y$，模型通过跨语言映射将两种语言的单语词嵌入映射至同一空间中对齐。跨语言映射步骤主要基于文献[27]提出的跨语言映射框架VecMap实现，分为归一化、白化、正交映射、去白化、重赋权重5个步骤。框架使用汉语与越南语的单语词嵌入空间$\boldsymbol{X}$、$\boldsymbol{Y}$作为输入，最终得到汉语与越南语的共享词嵌入空间，使具有相同含义的汉越单语词嵌入在空间中彼此接近。

### 1.4　汉越词典归纳

双语词典归纳(Bilingual Lexicon Induction，BLI)是评测跨语言词嵌入对齐准确性的通用任务，对于词典外任何一个未经标注的汉语或越南语单词，均可根据空间余弦相似度来查找该词在共享空间中对应的翻译词。设$w_i^X$与$w_j^Y$分别为汉语与越南语单词，$E_i^X$、$E_j^Y$分别为两个单词在共享空间中对应的汉语与越南语词嵌入，两词余弦相似度的计算过程如式(6)所示：

$$\cos(w_i^X,w_j^Y)=\frac{E_i^X E_j^Y}{\left\| E_i^X \right\|_2 \times \left\| E_j^Y \right\|_2} \tag{6}$$

其中：$\|*\|_2$代表2-范数。

例如在汉语到越南语的正向词典归纳任务中，以汉语单词“耳朵”作为检索词，经过计算选取余弦相似度最高的越南语单词“tai”作为候选词，构建汉越对齐词对“耳朵，tai”。通过词典归纳任务，可以为所有汉越单语词嵌入查找对应的翻译词，实现汉越跨语言词嵌入。

基于以上思想，本文提出融合词簇约束的汉越跨语言词嵌入模型，由汉越单语词嵌入训练、词簇对齐数据融合、映射矩阵训练和跨语言映射四部分组成。模型首先使用独立的单语语料训练获取汉越单语词嵌入，然后基于双语词典中的对齐词对构建词簇对齐数据融入映射矩阵的训练过程中，最后通过跨语言映射获取汉越共享词嵌入空间，模型架构如图2所示。

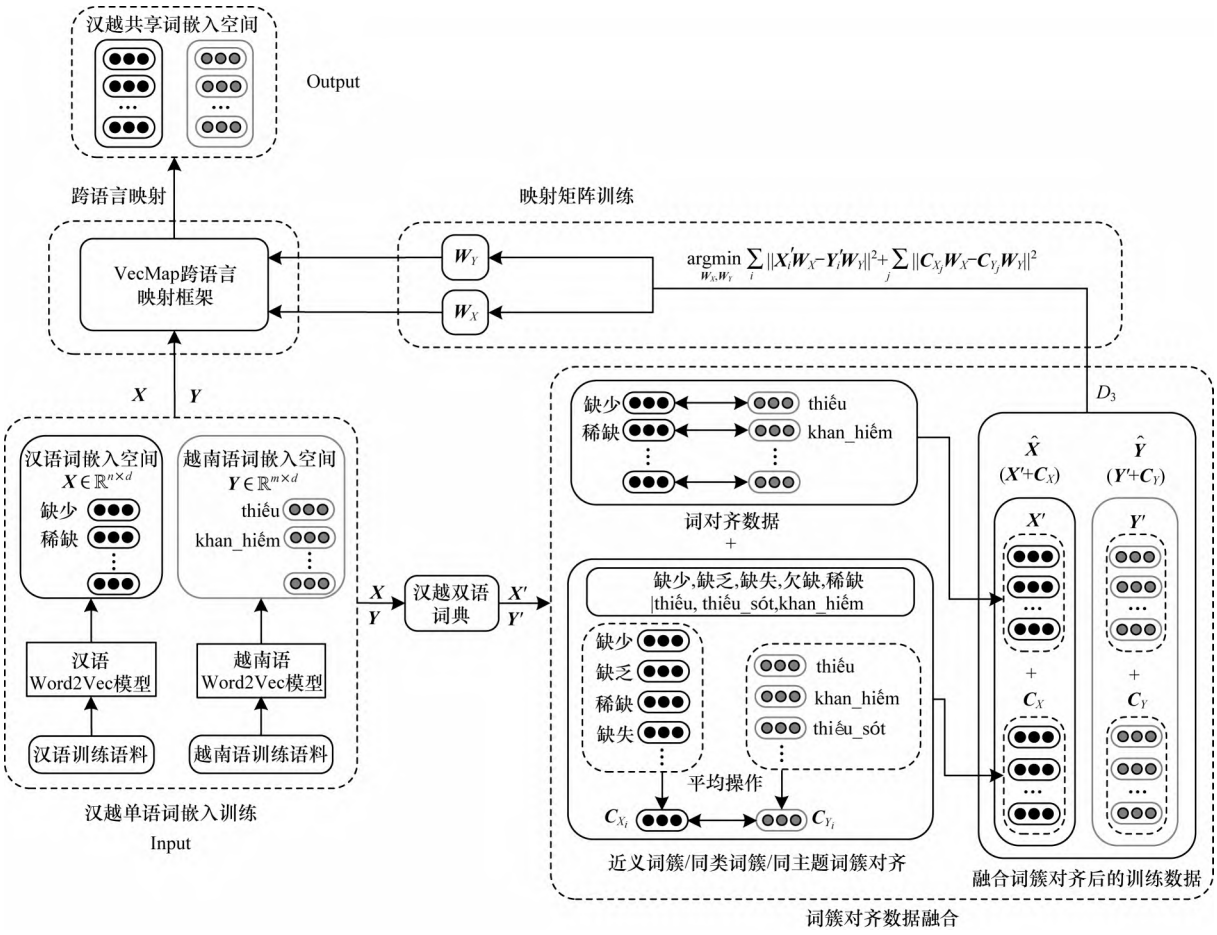



**图2 融合词簇约束的汉越跨语言词嵌入模型**

**Fig.2 Chinese-Vietnamese cross-lingual word embedding model with word cluster constraints**

## 2 实验数据准备及实验设置

### 2.1 汉越单语词嵌入训练

汉语与越南语均使用开源新闻数据集作为单语训练语料,汉语的单语语料来源于brightmart新闻数据集,使用jieba工具进行分词。越南语单语语料来源于binhvq新闻数据集,使用Vncorenlp工具进行分词。

汉语与越南语的单语词嵌入训练采用相同的参数设置,均使用Word2Vec模型中的CBOW(Continuous Bag-Of-Words)模型进行训练,词嵌入维度为300维,词窗大小设置为10,最低词频为50,迭代次数为5轮。经过训练,共得到32万汉语词嵌入和15万越南语词嵌入。

### 2.2 汉越双语词典构建

有监督跨语言词嵌入模型的效果很大程度上依赖于双语词典的质量,本文使用同主题词对和高频词对两种数据构建高质量的汉越双语词典。词典中的同主题词对来源于多语公开数据集CLDR(Unicode Common Locale Data Repository),其中包含月份、数字、颜色等常用主题词的对齐词对,如表1所示。这些词属于日常生活中的常用词,具有较高的词频和丰富的语义资源。

**表1 同主题对齐词对**

**Table 1 Aligned word pairs with the same subject**

| 主题 | 对齐词对 |
|---|---|
| 月份 | 一月,tháng_một;二月,tháng_hai;… |
| 数字 | 一,một;二,hai;三,ba;… |
| 颜色 | 红色,đỏ;粉红色,hồng;… |
| ⋮ | ⋮ |

传统双语词典依据源语言训练语料的词频构建,该策略认为高频词具有更高的权重和更丰富的语义特征。但汉语不同于其他语言,词频最高的词往往是一些助词、介词或单个文字,例如“的”、“在”、“了”等。这些词并不具备具体的含义,很难使映射矩阵捕捉到较为精确的映射关系。针对这一问题,本文提出一种面向汉越场景的高频词对构建流程。首先利用网上开源资源构建停用词表,剔除汉语高频词中的助词、介词和语气词,然后对剩余高频词进行人工筛选,去除特有名词及噪声词,最后使用Lingea在线词典人工标注相应的越南语翻译,并剔除不包含在汉越单语词嵌入文件中的OOV(Out Of Vocabulary)词对。本文一共构建了5 500对高质量的汉越双语词对,并从中随机抽取500对作为测试词典Test,剩余5 000对作为汉越双语词典Seed。

### 2.3　词簇对齐数据的构建

为使映射矩阵能够更好地学习到相近词间的共性特征及映射关系,词典中的词簇对齐数据主要利用近义词和同类词两种关系进行构建,构建流程如图3中1)、2)所示。首先查找汉越双语词典中越南语翻译相同的词对,这些词对通常具有极为相近的含义,可以构建为一个初始的词簇对齐。然后使用在线词典和开源词库Babelnet查找初始词簇在词典中的近义词对和同类词对用于扩充词簇对齐。除近义词簇和同类词簇外,词典中来源于多语公开数据集CLDR的同主题词对作为一种封闭词类,本身具有一定的弱相关性,可以构建为大范围的隐性对齐词簇,如图3中3)所示。3种不同类型的词簇对齐数据共同构成词簇对齐词典Seed_Cls,各词典规模如表2所示。

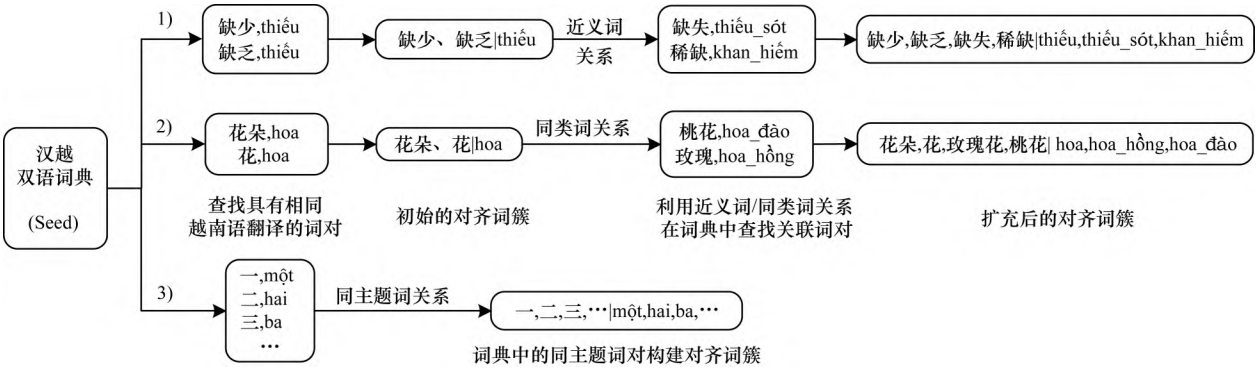



**图3　对齐词簇构建流程**

**Fig.3　Alignment word clusters construction process**

**表2　词典规模**

**Table 2　The scale of dictionaries**

| 词典 | 规模数/对 |
|---|---|
| 汉越双语词典 | 5 000 |
| 词簇对齐词典 | 829 |
| 测试词典 | 500 |

### 2.4　跨语言映射参数设置

为便于同基线模型进行对比,本文在单语词嵌入上采用相同的维度设置,模型中的汉语与越南语词嵌入维度均为300维。由于簇心嵌入是经单语词嵌入进行平均操作获得,因此其维度也为300维。此外,训练得到的映射矩阵 $\boldsymbol{W}_X$ 与 $\boldsymbol{W}_Y$ 的大小为300×300维,同词嵌入维度设置保持一致。VecMap跨语言映射框架中的归一化步骤使用长度归一化(unit)和中心化(center)作为预处理,执行顺序为[unit,center,unit]。框架其余步骤中的参数设置均同文献[27]保持一致,白化步骤中模型使用ZCA白化;在重赋权重步骤中,汉语与越南语的权重值分别设置为0.5和0.5。

### 2.5　评价指标

为更好地与现有工作进行比较,本文采用同mikolov、Artetxe、Conneau等在词典归纳任务上相同的评价指标,以词汇对齐的准确率P@$N$(选取$N$个候选词时的对齐准确率)作为衡量模型效果的标准,具体计算过程如式(7)所示:

$$P_{\text{准确率}}=\frac{\sum_{i=1}^{T}\left\|C\left(w_i\right)\right\|}{T}\times 100\% \tag{7}$$

其中:$T$代表测试词典的规模;$C\left(w_i\right)$代表模型依据余弦相似度为单词$w_i$检索到的$N$个候选词集合,若集合中包含正确的翻译词则取1,否则取0。

## 3　实验结果与分析

### 3.1　汉语到越南语词典归纳任务评测

为验证融合词簇对齐方法的有效性,模型同3个有监督模型和1个无监督模型进行了对比,基线模型设置如下:

1)Multi_w2v模型。文献[19]基于线性回归思想提出的跨语言Word2Vec模型,使用随机梯度下降最小化双语词典词对间的均方误差(Mean Squared Error,MSE)来学习映射矩阵。

2)Orthogonal模型。文献[21]提出的正交映射模型引入了长度归一化与中心化处理,并为映射矩阵添加正交约束。

3)VecMap模型。文献[27]提出的双向正交映射模型分别为源语言和目标语言训练单独的正交映射矩阵,并将两种语言的词嵌入映射至同一共享空间。

4)Muse模型。文献[11]基于无监督思想,利用单语嵌入空间之间的相似性,使用对抗的方式学习映射矩阵。

所有模型使用相同数据集进行训练与测试,参数设置同各文献保持一致,并在汉语到越南语的正向词典归纳任务上进行了对比,实验结果如表3所示。

**表3　汉越正向词典归纳任务实验结果**

**Table 3　Experimental results of the Chinese-Vietnamese forward lexicon induction tasks**　%

| 模型 | P@1 | P@5 |
|---|---|---|
| Muse | 37.00 | 62.60 |
| Multi_w2v | 42.00 | 67.00 |
| Orthogonal | 41.00 | 65.80 |
| VecMap | 42.80 | 69.20 |
| ClsMap | 45.00 | 71.40 |

分析表3的实验数据可知,融合词簇对齐约束的方法可以有效提升汉越低资源场景下跨语言词嵌入的对齐准确率,模型效果明显优于其他传统方法。由于汉越语言差异性大,词嵌入空间相似度低,Muse无监督模型在实验中的表现明显弱于其他有监督基线模型。而在有监督模型的对比中,基于线性回归方法实现的Multi_w2v模型在P@1和P@5任务上的效果优于单向正交映射模型Orthogonal,但表现不如基于双向正交映射的VecMap模型。本文模型ClsMap相较于表现最好的基线模型,在P@1和P@5任务上的对齐准确率提升了2.2个百分点。实验结果充分证明了在训练过程中融合词簇对齐约束的方法可以有效提升汉越低资源场景下映射矩阵在非标注词上的泛化性,提高了汉越双语空间的对齐准确率。

### 3.2 越南语到汉语词典归纳任务评测

为验证融合词簇对齐的方法在反向词典归纳任务上的表现,本文将训练集与测试集中的语言进行了置换,以越南语作为源语言,汉语作为目标语言进行了测试,实验结果如表4所示。

**表4 汉越反向词典归纳任务实验结果**

**Table 4 Experimental results of the Chinese-Vietnamese reverse lexicon induction tasks** %

| 模型 | P@1 | P@5 |
|---|---|---|
| Muse | 30.66 | 53.07 |
| Multi_w2v | 26.00 | 48.60 |
| Orthogonal | 29.39 | 53.28 |
| VecMap | 28.54 | 53.28 |
| ClsMap | 28.33 | 54.76 |

分析表4可知,在以越南语作为检索词的反向词典归纳任务中,本文方法在P@5任务上的对齐准确率达到了54.76%,相比VecMap模型与Orthogonal模型提升了1.48个百分点,取得了最好效果。而在P@1任务上,无监督方法取得了最好效果,但除Multi_w2v模型外,有监督模型与无监督模型在P@1任务上的差异并没有汉越正向词典归纳任务中那样明显。主要原因是由于双语词典是以汉语为源语言进行构建所致,这一问题在初始词簇的构建中就有所体现,词典中包含许多越南语相同的词对,证明汉语相较于越南语的标注更为精确。因此,用于训练的汉越双语词典在反向词典归纳任务中并不具备优势。ClsMap模型虽然在P@1任务上的表现一般,但在多候选词的P@5任务上依然取得了最好效果,证明融合词簇对齐约束的方法在越南语到汉语的反向词典归纳任务中仍具有一定的提升效果。

### 3.3 不同词典规模对模型效果的影响

为验证本文方法在低资源场景下的有效性,模型在不同规模的汉越双语词典上进行了对比实验。词典以2 000词作为最低规模,并逐次扩展到500词。模型在不同词典规模下的实验结果如表5所示,其中,词簇词典规模代表从双语词典中提取出的对齐词簇数量。

**表5 本文模型在不同词典规模下的对齐准确率**

**Table 5 Alignment accuracy of this model under different dictionary scales**

| 词典规模数/对 | 词簇词典规模数/对 | Zh→Vi/% | | Vi→Zh/% | |
|---|---|---|---|---|---|
| | | P@1 | P@5 | P@1 | P@5 |
| 2 000 | 440 | 43.00 | 67.00 | 21.56 | 51.80 |
| 2 500 | 556 | 43.80 | 67.20 | 22.83 | 53.28 |
| 3 000 | 648 | 45.20 | 69.20 | 24.95 | 54.55 |
| 3 500 | 710 | 45.40 | 71.20 | 25.58 | 54.33 |
| 4 000 | 758 | 43.80 | 71.00 | 26.64 | 54.55 |
| 4 500 | 807 | 44.60 | 72.00 | 26.00 | 54.76 |
| 5 000 | 829 | 45.00 | 71.40 | 28.33 | 54.76 |

通过对比表5与表3可知,在汉语到越南语的正向词典归纳任务(Zh→Vi)中,融合词簇对齐约束的方法仅使用2 000词就超越了大部分基线模型在5 000词上的训练效果。当词典规模达到3 000词时,通过融合从中提取到710条词簇对齐信息,模型在P@1和P@5任务上的对齐效果就已超越了最好基线模型在5 000词上的训练效果。随着词典规模的增长,在4 000~5 000词时,模型在P@5任务上的效果逐步趋于稳定,但在P@1任务上的效果有略微下降,推测这是由于在词典扩展过程中引入了更多近义词所导致,例如“好看,đẹp”与“漂亮,xinh”。这种近义词对在共享词嵌入空间中的距离较近,容易使映射出现偏差,导致模型在单个候选词任务上的效果下降。但随着数据规模的扩展,模型效果再次提升。在越南语到汉语的反向词典归纳任务(Vi→Zh)中,随着词典规模的扩大,模型在P@1和P@5任务上的效果稳步提升,波动并不明显,并在5 000词规模时取得了最好效果。

实验结果表明,相比传统跨语言词嵌入方法,融合词簇对齐约束的方法可以使模型在有限规模的双语词典中学习到更为精确的映射关系,以提升汉越双语空间的对齐效果,验证了本文方法在汉越低资源任务上的有效性。

### 3.4 词簇类型及其组合对模型效果的影响

为探索不同类型的词簇对齐信息及其组合对模型效果的影响,本文依据构建时所使用的关联关系将词簇词典划分为近义词簇、同类词簇和同主题词簇三部分。其中,近义词簇251条、同类词簇534条、同主题词簇44条,模型在不同类型词簇及其组合上的实验结果如表6所示。

表6　本文模型在不同类型词簇及其组合下的对齐准确率

Table 6　Alignment accuracy of this model under different types of word clusters and their combinations

| 模型 | 词簇词典组合 | 词簇词典规模数/对 | P@1/% | P@5/% |
|---|---|---|---|---|
| VecMap | 无 | 0 | 42.80 | 69.20 |
| ClsMap | 近义词簇 | 251 | 44.60 | 71.00 |
| | 同类词簇 | 534 | 43.60 | 70.60 |
| | 同主题词簇 | 44 | 44.20 | 69.60 |
| | 近义+同类词簇 | 785 | 45.00 | 71.20 |
| | 近义+同主题词簇 | 295 | 44.60 | 71.40 |
| | 同类+同主题词簇 | 578 | 43.60 | 71.20 |
| | 近义+同类+同主题词簇 | 829 | 45.00 | 71.40 |

通过对比表6中基线模型与本文模型在汉越正向词典归纳任务中的实验结果可知,融合不同类型的词簇对齐信息均可提升模型在P@1与P@5任务上的对齐准确性。在前三组单一类型词簇的融合实验中,近义词簇的效果明显优于其他两种词簇,这是因为近义词通常具有极为相近的含义,在单语嵌入空间中的距离更近,能使映射矩阵较为精准地学习到汉越相近语义词间共有的结构特征和映射关系,提升模型在非标注词上的泛化能力。但由于越南语的近义词较难获取,导致词典提取到的词簇数量有限,模型难以达到最优效果,因此还需引入其他类型的对齐词簇作为补充。而同类词簇对语义相关性的要求较低且容易获取,可以作为近义词簇的补充融入训练。从实验数据可知,模型在近义词簇与同类词簇组合上的训练效果已十分接近最优。虽然使用同类词簇学习到的结构特征不如近义词簇精确,但依赖其数量上的优势依然可以给模型带来较大提升。除近义词簇与同类词簇外,词典中的同主题词对间本身具有一定的弱相关性,可以构建为一种大范围的对齐词簇。虽然同主题词簇因涵盖范围较广,但其主体来源于多语公开数据集CLDR,无需人工标注,且在同近义词簇与同类词簇进行组合后,依然可以给模型带来微小的提升,以取得最佳的对齐效果。

### 3.5　不同参数设置对模型效果的影响

为更好地与现有工作进行比较,模型采用同mikolov、Artetxe、Conneau等相同的词嵌入维度设置。因此,本文将重点探索归一化步骤设置及重赋权重步骤中汉语与越南语的权重值配比对模型效果的影响。

1)在归一化设置中,鉴于文献[21]已详细分析了归一化对跨语言词嵌入模型效果的影响,本文直接使用其推荐的归一化方式作为组合的基础选项,即长度归一化(unit)和中心化(center)。由于连续执行长度归一化(unit+unit)和中心化(center+center)的操作是无意义的,因此共设置了如下6组归一化组合,模型在不同组合上的实验效果如表7所示。

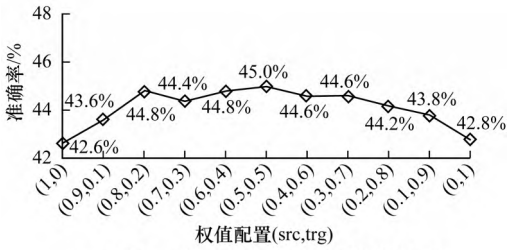
表7　不同归一化组合对模型效果的影响

Table 7　The influence of different normalization combinations on the model effect　%

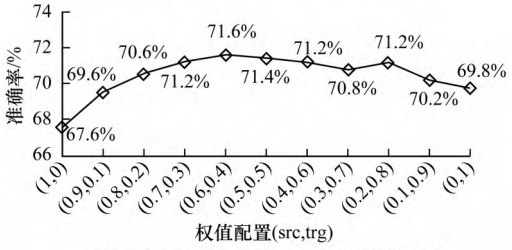
| 归一化步骤组合 | P@1 | P@5 |
|---|---|---|
| unit | 44.20 | 71.80 |
| center | 44.00 | 71.40 |
| unit+center | 44.80 | 71.40 |
| center+unit | 44.60 | 71.80 |
| unit+center+unit | 45.00 | 71.40 |
| center+unit+center | 44.60 | 72.00 |

通过分析表7中的实验数据可知,使用[center+unit+center]作为预处理组合时,模型在汉越正向词典归纳任务P@1与P@5上取得了较好的效果,但考虑到跨语言词嵌入在实际应用场景中的主要目标是实现双语词间的精确对齐,因此最终选取在P@1任务上表现最好的预处理组合[unit+center+unit]作为首选的归一化设置。

2)考虑到不同语言的单语训练语料有时更偏向于某一特定领域,如金融、法律等,此时依据语料词频构建的双语词典无法准确代表单语词嵌入的整体分布,导致学习到的映射关系存在一定的偏差。因此,提出重赋权重步骤,通过为两种语言赋予不同比例的权重值来对映射后的词嵌入进行微调,以实现更好的对齐效果[27]。以汉越正向词典归纳P@1与P@5任务为例,模型在不同比例权重值上的对齐效果如图4所示。其中,src代表汉语对应的权重值,而trg代表越南语对应的权重值。




(a)不同权重值配比在P@1任务上的对齐效果




(b)不同权重值配比在P@5任务上的对齐效果

图4　模型在不同权重值比例上的对齐效果

Fig.4　Alignment effect of the model on different weight value scales

从图4中数据可知,当汉语与越南语的权重值配比分别为0.5和0.5时,模型在P@1任务上取得了最佳的对齐效果;当权重值配比为0.6和0.4时,模型在P@5任务上取得了最佳效果。此外,随着两端权重值配比

差异的增大,模型效果逐渐降低。考虑到P@1任务相较于P@5任务对齐难度更大,本文最终选取(0.5,0.5)作为重赋权重步骤中汉语与越南语的权值配比。

### 3.6 实例分析

为直观反映融合词簇约束方法对模型映射准确性的影响,本文选取了3个汉越词典归纳任务中的实例进行了对比说明,如表8、表9所示。

**表8 汉越词典归纳任务实例**

**Table 8 Examples of Chinese-Vietnamese lexicon induction tasks**

| 实例 | 检索词 | 模型 | 候选词 |
|---|---|---|---|
| 1 | 礼拜五 | VecMap | thứ,thứ_bảy,thứ_sáu,chủ_nhật,tuần<br>东西,星期六,星期五,星期日,星期 |
| | | ClsMap | thứ_bảy,thứ_sáu,chủ_nhật,thứ,tuần<br>星期六,星期五,星期日,东西,星期 |
| | | 对齐词簇 | [星期一,星期二,…\| thứ_hai,thứ_ba,…] |
| 2 | 情感 | VecMap | tình_cảm,tâm_hồn,nội_tâm,cảm_xúc,xúc_cảm<br>情事,灵魂,内心,情感,感觉 |
| | | ClsMap | cảm_xúc,tình_cảm,nội_tâm,tâm_hồn,xúc_cảm<br>情感,情事,内心,灵魂,感觉 |
| | | 对齐词簇 | 无 |

**表9 缺陷实例**

**Table 9 Example of defects**

| 实例 | 检索词 | 模型 | 候选词 |
|---|---|---|---|
| 3 | 紫色 | VecMap | đỏ,nâu,tím,hồng,đỏ_son<br>红色,棕色,紫色,粉红,朱红 |
| | | ClsMap | nâu,đỏ_son,đỏ,hồng,tím<br>棕色,朱红,红色,粉红,紫色 |
| | | 对齐词簇 | [红色,白色,黄色,…\| đỏ,trắng,vàng,…] |

表8为基线模型VecMap与本文模型ClsMap在汉越词典归纳任务中的两个实例,模型分别输出5个与检索词余弦相似度最高的越南语单词作为候选词,候选词下面为对应的汉语翻译。表8的对齐词簇代表词簇词典中是否包含同检索词相关的对齐词簇,例如实例1中以非标注词“礼拜五”作为检索词,同时词簇词典中也含有表示“星期”的同主题对齐词簇。通过分析实例1可知,在P@5任务上,本文模型中的正确翻译词“thứ_sáu”在相似度排序上相比基线模型前移了一位,效果更好。此外,基线模型的第一候选词为“thứ”,对应汉语为“东西”,同检索词的语义相关性较低,而ClsMap模型中的前3位候选词同检索词的语义相关性更高。实例1充分证明融合词簇对齐信息可以使映射矩阵学习到不同语言相近语义词间的共性特征及映射关系,能够通过映射拉近汉越相近词在共享空间中的距离,减小对齐难度。实例2则体现了当词簇词典中无相关词簇对齐信息时,模型在非标注词上的性能。通过对比表中数据可知,以非标注词“情感”作为检索词时,基线模型在P@1任务上的对齐并不准确,而本文方法在P@1任务上实现了精确对齐,同时候选词的相似度排序同检索词的语义相关性更强。实例2充分证明模型可以将学习到的相近词间的映射关系迁移到其他未标注词簇上,进一步提升模型在非标注词上的泛化能力,改善汉越低资源场景下双语空间的对齐效果。

然而,本文方法在个别实例上也体现出了一些问题,例如表9中的实例3所示,两个模型以非标注词“紫色”作为检索词,并且词簇词典中包含表示“颜色”的同主题对齐词簇。通过对比可知,在P@5任务上,本文方法相比基线模型,正确翻译词在相似度排序上后退了两位。推测这是由于在词簇对齐数据中,缺乏带有“紫色”含义的词,而具有“红色”含义的词较多。因此,融合词簇对齐约束后,结果中的“đỏ_son”、“đỏ”、“hồng”的距离更加接近,导致正确翻译词“tím”后移,但该实例也从侧面反映出本文方法能有效拉近相近语义词间的距离。

## 4 结束语

针对汉越低资源场景下语言差异性大、双语词典规模小导致跨语言词嵌入对齐效果较差的问题,本文提出一种融合词簇对齐约束的方法。通过使用近义词、同类词和同主题词3种类型的关联关系抽取汉越双语词典中的词簇对齐信息融入映射矩阵的训练中,使映射矩阵学习到不同语言相近词间的共性特征及映射关系,以提升模型在非标注词上的泛化性。实验结果表明,本文方法在汉越词典归纳任务中P@1和P@5上的对齐效果相比基线模型均有明显提升,能有效提高汉越低资源环境下双语空间的对齐准确性。由于依据双语词典提取出的对齐词簇虽然质量较高但数量有限,因此下一步考虑引入迭代的思想使模型自行构建高质量的对齐词簇用于训练,以提升模型的映射准确性。

## 参考文献

[1] MOGADALA A, RETTINGER A. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification[C]//Proceedings of 2016 International Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Washington D. C., USA: IEEE Press, 2016: 692-702.

[2] LIU P F, QIU X P, HUANG X J. Recurrent neural network for text classification with multi-task learning[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2016: 2873-2879.

[3] ESULI A, MOREO A, SEBASTIANI F. Funnelling: a new ensemble method for heterogeneous transfer learning and its application to polylingual text classification[J]. ACM Transactions on Information Systems, 2019, 37(3): 1-30.

[4] CHEN X L, SUN Y, ATHIWARATKUN B, et al. Adversarial deep averaging networks for cross-lingual sentiment classification[J]. Transactions of the Association for Computational Linguistics, 2018, 6: 557-570.

[5] SINGH P, LEFEVER E. Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings[C]//Proceedings of the 4th Workshop on Computational Approaches to Code Switching. Marseille, France: European Language Resources Association Press, 2020: 45-51.

[6] MADHYASTHA P S, ESPAÑA-BONET C. Learning bilingual projections of embeddings for vocabulary expansion in machine translation[C]//Proceedings of the 2nd Workshop on Representation Learning for Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2017: 139-145.

[7] 陈玺,杨雅婷,董瑞. 面向汉维机器翻译的BERT嵌入研究[J]. 计算机工程,2021,47(12):112-117.
CHEN X, YANG Y T, DONG R. Research on BERT embedding for Chinese-Uyghur machine translation[J]. Computer Engineering, 2021, 47(12): 112-117. (in Chinese)

[8] MULLOV C, PHAM N Q, WAIBEL A. Unsupervised transfer learning in multilingual neural machine translation with cross-lingual word embeddings[EB/OL]. [2021-10-20]. https://arxiv:2103.06689v1.

[9] TSAI C T, ROTH D. Cross-lingual wikification using multilingual embeddings[C]//Proceedings of 2016 International Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics, 2016: 589-598.

[10] PAN X M, GOWDA T, JI H, et al. Cross-lingual joint entity and word embedding to improve entity linking and parallel sentence mining[C]//Proceedings of the 2nd International Workshop on Deep Learning Approaches for Low-Resource NLP. Stroudsburg, USA: Association for Computational Linguistics, 2019: 56-66.

[11] LAMPLE G, CONNEAU A, et al. Word translation without parallel data[C]//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ICLR Press, 2018: 325-338.

[12] ALVAREZ-MELIS D, JAAKKOLA T. Gromov-Wasserstein alignment of word embedding spaces[C]//Proceedings of 2018 International Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2018: 1881-1890.

[13] LI Y L, ZHANG Y H, YU K, et al. Adversarial training with Wasserstein distance for learning cross-lingual word embeddings[J]. Transactions of the Applied Intelligence, 2021, 51(11): 7666-7678.

[14] ARTETXE M, LABAKA G, AGIRRE E. Learning bilingual word embeddings with (almost) no bilingual data[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2017: 451-462.

[15] MARCHISIO K V, KOEHN P, XIONG C H. An alignment-based approach to semi-supervised bilingual lexicon induction with small parallel corpora[C]//Proceedings of International Machine Translation Summit. [S. l.]: MT Summit Press, 2021: 293-304.

[16] ZHAO X, WANG Z H, WU H, et al. Semi-supervised bilingual lexicon induction with two-way interaction[C]// Proceedings of 2020 International Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2020: 2973-2984.

[17] PATRA B, MONIZ J R A, GARG S, et al. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 184-193.

[18] SØGAARD A, RUDER S, VULIĆ I. On the limitations of unsupervised bilingual dictionary induction[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 778-788.

[19] MIKOLOV T, LE Q V, SUTSKEVER I. Exploiting similarities among languages for machine translation[EB/OL]. [2021-10-20]. https://arxiv:1309.4168v1.

[20] XING C, WANG D, LIU C, et al. Normalized word embedding and orthogonal transform for bilingual word translation[C]// Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics, 2015: 1006-1011.

[21] ARTETXE M, LABAKA G, AGIRRE E. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance[C]//Proceedings of 2016 International Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2016: 2289-2294.

[22] DOVAL Y, CAMACHO-COLLADOS J, ESPINOSA-ANKE L, et al. Improving cross-lingual word embeddings by meeting in the middle[C]//Proceedings of 2018 International Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2018: 294-304.

[23] AZPIAZU I M, PERA M S. Hierarchical mapping for crosslingual word embedding alignment[J]. Transactions of the Association for Computational Linguistics, 2020, 32(8): 361-376.

[24] RUDER S, VULIĆ I, SØGAARD A. A survey of cross-lingual word embedding models[J]. Journal of Artificial Intelligence Research, 2019, 65: 569-631.

[25] HUANG L F, CHO K, ZHANG B L, et al. Multi-lingual common semantic space construction via cluster-consistent word embedding[C]//Proceedings of 2018 International Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2018: 250-260.

[26] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2013: 3111-3119.

[27] ARTETXE M, LABAKA G, AGIRRE E. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations[C]//Proceedings of the 32th AAAI Conference on Artificial Intelligen. New Orleans, USA: AAAI Press, 2018: 5012-5019.

编辑 索书志