文章编号：1003-0077(2021)07-0047-07

基于半监督的汉缅双语词典构建方法

毛存礼[1,2]，陆 杉[1,2]，王红斌[1,2]，余正涛[1,2]，吴 霞[1,2]，王振晗[1,2]

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

摘 要：汉缅双语词典是开展机器翻译、跨语言检索等研究的重要数据资源。当前在种子词典的基础上使用迭代自学习的方法在平行语料中抽取双语词典取得了较好的效果，然而针对低资源语言汉语—缅语的双语词典抽取任务，由于双语平行资源匮乏，基于迭代自学习的方法不能得到有效的双语词向量表示，致使双语词典抽取模型准确度较低。研究表明，可比语料中相似词语往往具有相似的上下文，为此，该文提出了一种基于半监督的汉缅双语词典构建方法，通过利用预训练语言模型来构建双语词汇的上下文特征向量，对基于可比语料和小规模种子词典的迭代自学习方法得到的汉缅双语词汇进行语义增强。实验结果表明，该文提出的方法相较于基线方法有明显的性能提升。

关键词：汉缅双语；种子词典；迭代自学习；预训练语言模型；上下文特征；半监督

中图分类号：TP391 **文献标识码**：A

Semi-supervised Chinese-Burmese Bilingual Dictionary Construction

MAO Cunli[1,2], LU Shan[1,2], WANG Hongbin[1,2], YU Zhengtao[1,2], WU Xia[1,2], WANG Zhenhan[1,2]

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

Abstract: Chinese-Burmese bilingual dictionary is an important data resource for research on machine translation and cross-language retrieval, etc. At present, the iterative self-learning method based on small-scale seed dictionary has achieved good results in extracting bilingual dictionaries from parallel corpus. However, for low-resource languages like Chinese-Burmese bilingual dictionary extraction task, due to the lack of bilingual parallel resources, the method based on iterative self-learning can not get effective bilingual word vector representation, resulting in the low accuracy of bilingual dictionary extraction model. Recent studies suggest that similar words in comparable corpora often have similar contexts. Therefore, this paper proposes a semi-supervised method for constructing Chinese- Burmese bilingual dictionary. By using the pre training language model, the context feature vector of bilingual vocabulary is constructed. The Chinese-Burmese bilingual vocabulary obtained by the iterative self-learning method of comparable corpus and small-scale seed dictionary is semantically enhanced. The experimental results show that the proposed method has a significant improvement comparing with the baseline method.

Keywords: Chinese-Burmese bilingual; seed dictionary; iterative self-learning; pre-trained language model; contextual feature; semi-supervised

0 引言

汉缅双语词典构建对开展该语言的自然语言处理相关研究具有重要的研究价值。但由于缅甸语是一种典型的资源稀缺型语言，其相关语料的采集、标注等工作具有一定的难度。缅甸语属于东南亚语言，与中文、英文不同，其语言的最小单位为音节，直

收稿日期：2020-10-26 **定稿日期**：2020-12-07

基金项目：国家自然科学基金(61732005,61662041,61761026, 61866019,61972186)；云南省应用基础研究计划重点项目(2019FA023)；云南省中青年学术和技术带头人后备人才项目(2019HB006)

接采用人工的方式对缅甸语进行双语词典的标注，需要标注人员对双语语言知识极为了解并且时间成本极高。

利用词对齐[1]等统计的方法直接构建汉缅双语词典是一种有效的方式，但是，从互联网中很难直接获得大规模的汉缅平行语料，汉缅可比语料相比于平行句子更容易获得，因此如何利用相对丰富的双语可比语料来构建汉缅双语词典，成为汉缅双语词典构建任务中亟待解决的问题。

由于汉语和缅语的词典规模受限，直接利用双语词典作为弱监督信号来学习公共语义空间中的双语映射关系[2]，会导致学习到的双语词嵌入的准确率较低。又因训练词向量的双语语料不够充分，会出现如中文语料中“工人”的上下文信息和缅甸语语料中“老师”的上下文信息比较相似的情况，这导致汉语意思为“工人”的词映射会和缅甸语“ဆရာ(老师)”更近，却距离正确的缅甸语词映射“အလုပ်သမား(工人)”更远。以上现象会带来翻译不准确、语义空间不匹配的问题。

针对以上的问题，本文提出一种基于半监督学习的汉缅双语词典构建方法，使用汉缅可比语料和小规模的种子词典基于迭代自学习的方法，学习汉-缅双语词映射关系，根据映射关系得到汉缅双语候选词集合，再使用多语言 BERT[3]得到这些词语集合的上下文特征作为约束，抽取得到质量更高的汉缅双语词典。本文的主要贡献如下：

(1) 针对于缅甸语这种低资源语言，利用汉缅可比语料基于迭代自学习的方法扩充了汉缅双语词典，解决了由于汉缅平行数据稀缺导致抽取的双语词典不准确的问题。

(2) 基于 BERT 利用汉缅双语文档中的上下文信息约束双语词映射关系，进一步提高了抽取得到的汉缅双语词典的质量。

(3) 在本文构建的汉缅可比文档数据集上，利用本文方法构建了规模大小为 3000 对的汉缅双语词典。

本文的第 1 节介绍了双语词典构建的相关工作；第 2 节介绍了基于半监督的汉缅双语词典构建方法；第 3 节通过实验对比证明本文方法的优势；第 4 节总结全文并指出下一步的研究方向。

1 相关工作

本文按照语料库资源的不同，将语料库构建的相关工作主要分为四类：①基于篇章级对齐语料的方法，②基于句子级对齐语料的方法，③基于种子词典的方法，④基于无监督的方法。

篇章级对齐语料是互联网资源中常见的可比语料，这类方法以依赖篇章级对齐语料为资源来抽取双语词典。Vulic 等人[4]利用 skip-gram 模型学习双语词嵌入，并用来抽取双语词典。

基于句子级对齐语料的方法，是以平行句对作为资源，利用句子中的对齐信息抽取双语词典。Chandar 等人[5]基于自动编码器学习双语词向量表示，以提高源句和目标句向量之间的相关性。Gouws 等人[6]提出了无对齐的双语词袋模型，学习单词的双语分布式表示形式。

基于种子词典的方法，通过现有的双语词典来学习双语词典映射关系，并通过该映射关系来构建双语词典。Wick 等人[7]利用一小部分人为提供的单词翻译工具，并在模型的目标函数中将单词翻译编码作为硬约束，从而抽取双语词典。Duong 等人[8]认为前人的工作中存在难以合并单语数据或无法处理一词多义问题，因此他们利用 EM 训练算法中的高覆盖字典来构建双语词典。Cao 等人[9]提出了一种分布模型，利用单语数据学习双语词嵌入，将学习得到的双语词嵌入在共享向量空间中对齐，将每个单词与其对应翻译进行组合，以此实现双语词典的构建。

基于无监督的方法，不需要平行语料和种子词典等双语监督信号，直接实现双语词典的抽取[10-14]。但是，由于汉语和缅语之间的语言差异性较大，直接通过无监督方法构建汉-缅双语词典效果有待提高。

针对缅甸语的语料构建工作主要如下：

AUNG 等人[15]利用中文句子中的实体以及实体类别、位置、长度等特征作为约束，提出一种基于汉-缅双语可比语料的双语实体抽取方法。Thu 等人[16]从英语语料库中筛选出两万句进行人工翻译，得到两万条英缅平行语料，并采用人工标注的方式对翻译得到的缅语语料进行分词、词性标记等工作。构建缅甸语语料库不单单是对词性、实体等内容的标注，还包括构建汉缅平行句对、双语词典等工作。这些工作目前很少有人关注，因为在汉缅语料收集、标注方面均存在一定的困难。

综上所述，针对汉缅双语词典的构建任务而言，汉缅双语平行语料稀缺且获取成本高昂。同时，基于种子词典的方法受限于低资源语言汉缅种子词典的规模及质量，其模型容易陷入局部最优，使得构建效果不太理想。基于对抗网络的方法，在两种语言

的差异性较小时，其构建的双语词典准确率较高，但在语言差异性较大时词典准确率不高。因此，本文减少了对大量平行语料和双语词典的需求，以更小的种子词典学习到更好的双语映射，并使用候选词的上下文特征作为约束，有效地从可比文档中抽取双语词典。

2　基于半监督的汉缅双语词典构建方法

我们提出了基于半监督学习的汉缅双语词典抽取方法，模型框架如图 1 所示，基本过程如下：

（1）使用 Word2Vec[17] 处理汉缅双语可比文档得到汉-缅单语的词向量表示。

（2）通过(1)中得到汉缅单语词向量，结合小规模种子词典作为枢轴，使用迭代自学习的方法，不断迭代扩充词典大小来学习得到更好的双语词向量映射关系，直至迭代收敛。

（3）从(2)中得到双语词向量中随机选取汉语词向量，根据余弦相似度找到 $N(N=1,5,10)$ 个相似度最高的缅语词向量，得到其候选词集合。

（4）通过设置窗口大小 $k(k=1,2,3,4)$，从可比文档中得到候选词和其前后各 k 个词所组成的上下文词组，结合多语言 BERT 得到候选词上下文词组的向量表示。

（5）计算出目标词汇上下文词组向量表示与候选词的上下文词组向量表示的余弦相似度，通过权重值加权得到最后的相似度，选取相似度最大的双语词对作为最后的抽取结果。

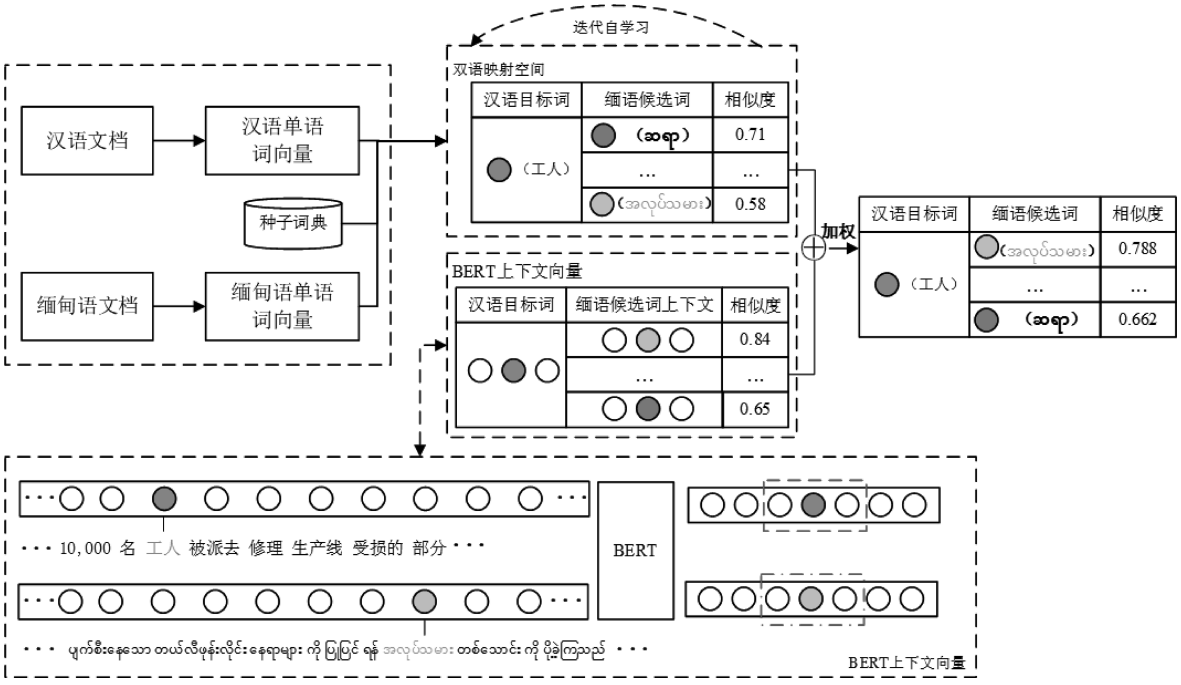

图 1　基于半监督学习的汉缅双语词典抽取框架

2.1　基于小规模种子词典的汉缅词嵌入映射学习

采用基于种子词典的方法学习汉缅双语词向量的映射矩阵，需要先使用一个小规模的双语种子词典，并通过该种子词典学习两种语言词向量间映射关系。

如图 1 所示，我们将现有的汉缅可比文档经过 Word2Vec 预处理后，训练生成汉语、缅语的单语词向量。$\boldsymbol{X}$ 表示汉语语料训练得到的词嵌入矩阵，$\boldsymbol{Z}$ 表示缅语语料训练得到的词嵌入矩阵。词嵌入矩阵的行数表示词的个数，列数表示词嵌入的维度，X_{i*} 表示汉语词嵌入矩阵 $\boldsymbol{X}$ 中第 i 个词的词嵌入，Z_{j*} 表示缅语词嵌入矩阵 $\boldsymbol{Z}$ 中第 j 个词的词嵌入。将种子词典表示为一个二维矩阵 $\boldsymbol{D}$，$\boldsymbol{D}_{ij}$ 为 1 时代表目标语言中的第 j 个单词是源语言中第 i 个单词的翻译。然后找到最佳映射矩阵 $\boldsymbol{W}^*$（$\boldsymbol{W}^*$ 为映射矩阵 $\boldsymbol{W}$ 的最优结果）让汉语词向量跟缅语词向量分布在同一个向量空间，使得映射源嵌入 $\boldsymbol{X}_{i*}\boldsymbol{W}$ 与目标嵌入 $\boldsymbol{Z}_{j*}$ 之间的欧几里德距离的平方和最小，映射矩阵 $\boldsymbol{W}^*$ 定义如式(1)所示。

$$\boldsymbol{W}^* = \arg\min_{W} \sum_i \sum_j \boldsymbol{D}_{ij} \| \boldsymbol{X}_{i*}\boldsymbol{W} - \boldsymbol{Z}_{j*} \|^2 \quad (1)$$

进一步，对每个单词的词向量做归一化处理，再对词向量的每一列取均值，最后再进行一次归一化处

理,同时将 $\boldsymbol{W}$ 约束为正交矩阵,其用于强制汉、缅单语不变性,防止单语性能的降低,同时产生更好的双语映射。在这种正交性约束下,最小化欧氏距离平方等于最大化点积,因此映射矩阵 $\boldsymbol{W}^*$ 被定义为式(2):

$$W^* = \underset{W}{\operatorname{argmax}} \operatorname{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{Z}^T\boldsymbol{D}^T) \tag{2}$$

$\operatorname{Tr}(\cdot)$ 表示所有主对角线上所有元素的和,通过计算 $W^* = \boldsymbol{U}\boldsymbol{V}^T$,得到最优正交解,$X^T\boldsymbol{D}\boldsymbol{Z} = \boldsymbol{U}\sum\boldsymbol{V}^T$ 是 $\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{Z}$ 的奇异值分解。由于字典矩阵 $\boldsymbol{D}$ 是稀疏矩阵,这可以有效地在线性时间内对汉缅字典条目数进行计算。

在模型迭代过程中,主要实现两个任务:一是通过种子词典,学习到最佳的双语映射关系矩阵 $\boldsymbol{W}^*$。二是通过上一步的映射关系,计算到最佳字典矩阵 $\boldsymbol{D}$。

2.2 基于迭代自学习的候选词选取

基于 2.1 节获得最佳映射矩阵 $\boldsymbol{W}^*$ 后,对于词典外任何一个没有对齐的单词,都可以根据映射后的空间余弦相似度来进行词对齐。

$$\cos_{\text{dic}}(w_i^S, w_j^T) = \frac{(W^* x_i)^T (z_j)}{\|W^* x_i\|_2 \times \|z_j\|_2} \tag{3}$$

设 S 表示汉语句子,T 表示缅语句子,通过式(3)计算汉语句子 S 中任意一个汉语词 w_i^S 与缅语句子 T 中所有缅语词 $w_j^T(j = 1,2,\cdots,n)$ 的相似度,其中 n 为缅语句子中词语数量,$\boldsymbol{x}_i$,$\boldsymbol{z}_j$ 分别为汉语词 w_i^S 与缅语词 w_j^T 的词向量表示,$\|\cdot\|_2$ 表示 2-范数。

通过计算,选取相似度最高的前 N 个缅语单词作为此汉语词相对应的缅语平行词的候选词集合,如图 1 所示,得到汉语词“工人”的 N 个缅语候选词以及对应的相似度得分。

2.3 基于 BERT 的候选词汇的上下文表征

传统词向量都属于静态词向量,其每个词汇的向量表示是唯一的,不会随着上下文的不同而改变,而通过 BERT 得到的上下文词组表示能充分地表达出上下文信息。因此本文计算汉缅双语候选词汇上下文特征向量的相似性来约束双语词典抽取过程。

多语言 BERT 是一种考虑词语位置关系及上下文语义的模型。基于此,本文采用 Google 开源的多语言 BERT① 模型来构造候选词汇的上下文特征表示,从候选词汇的前后单词中获得其上下文特征。BERT 对每个输入使用注意力机制,以获得当前输入与上下文语义的关系和自身所包含的信息。通过多层累加和多头注意力机制,不断获取当前输入更为合适的向量表示。所以利用多语言 BERT 模型训练双语词汇能得到更好的上下文特征表示。本文为避免远距离上下文信息对词本身表示造成干扰,使用的是窗口大小的上下文信息作为约束条件,而窗口大小是以词为单位计算的。考虑到缅甸语语言特性,在使用多语言 BERT 进行上下文信息编码时,不是像传统做法那样直接使用上下文句子作 Wordpiece,而是用分好词的上下文句子,对每个词语进行 Wordpiece 后把所有子词送入 BERT 模型得到编码,这样做是因为传统做法会出现把两个词切分在一个子词里的情况。此外使用 BERT 编码上下文信息时,不是像传统做法那样使用十二层输出的最后一层,而是使用最后四层求和作为最终的上下文向量,以确保上下文向量能包含最丰富的语义信息[3]。设 S_k^i 为在窗口大小为 k 的情况下,汉语句子 S 中第 i 个汉语词的上下文特征表示,T_k^j 为在窗口大小为 k 的情况下,缅语句子 T 中第 j 个缅语候选词的上下文特征表示,则汉语词 w_i^S 和缅语词 w_j^T 之间的上下文相似度计算如式(4)所示。

$$\cos_{\text{con}}(w_i^S, w_j^T) = \frac{(S_k^i)^T (T_k^j)}{\|S_k^i\|_2 \times \|T_k^j\|_2} \tag{4}$$

如图 1 所示,当窗口大小设置为 1 时,可比文档中汉语单词“工人”的上下文组合为“万名工人被派去”,而缅语“အလုပ်သမား(工人)”的上下文组合为“ရန်(修理)အလုပ်သမား(工人) တစ်သောင်း(万名)”,可以看出,汉缅语法结构非常不同。

2.4 基于上下文特征约束的汉缅双语词典抽取

基于种子词典的双语词典抽取方法通过小规模的种子词典学习到双语词汇间的映射关系,得到候选词汇,而候选词汇的上下文特征可以计算两个词在跨语言上下文上的分布相似性。将这两种方式结合起来,可以利用全局知识和上下文知识来计算相似度,使双语词汇的抽取更加可靠和准确。为此,我们将汉语词 w_i^S 和缅语词 w_j^T 通过 2.2 节中式(3)计算出全局相似度得分 $\cos_{\text{dic}}(w_i^S, w_j^T)$,通过 2.3 节中式(4)计算得到上下文相似度得分 $\cos_{\text{con}}(w_i^S, w_j^T)$,并把两个相似度得分进行线性组合。该组合是计算汉缅双语词汇与上下文双语词汇的综合相似性得分,定义如式(5)所示。

$$\begin{aligned}\text{Sim}_{\text{comb}}(w_i^S, w_j^T) = {} & \lambda \cos_{\text{con}}(w_i^S, w_j^T) \\ & + (1-\lambda)\cos_{\text{dic}}(w_i^S, w_j^T)\end{aligned} \tag{5}$$

① https://github.com/google-research/bert

其中，λ 是两种方法线性结合过程中的超参数，首先使用种子词典方法为汉语单词生成一个包含 N 个缅语候选词的列表，然后通过上下文特征向量计算候选列表词的相似度。最后，我们进行组合约束，组合过程是一次对基于种子词典抽取的候选词的重新排序，最终实现汉缅双语词汇的抽取。

3　实验

3.1　数据集及评价指标

由于目前还没有用于汉缅双语词典抽取的公开数据集，本文实验数据的采集首先是利用网络爬虫从维基百科[①]获取汉语、缅语相对应词条的篇章对齐文档。例如，在中文维基百科搜索“太阳”后的百科内容，和在缅甸维基百科中搜索“ဂြိုဟ်(太阳)”后的百科内容来作为篇章对齐文档，词条文章涉及政治、教育、文化、经济等领域。然后经过人工筛选句子数在 10 句至 30 句以内的对齐文档 600 篇，再经过数据去杂、分词等预处理后，文档结构如表 1 所示。

表 1　实验数据集

语言种类	汉语	缅语
可比文档数	600	600
句子总数	10 213	13 213
词总数	98 778	123 226

本文使用的种子词典通过网络爬取，然后人工校对，剔除字符长度大于 6 的中文词语，最后选取 5000 对双语互译词汇，其中词语类型分为名词、动词、形容词和代词，分别占比 95.64%、2.4%、1.74%、0.22%，部分示例如表 2 所示。

表 2　汉缅双语词示例

汉语	缅甸语	汉语	缅甸语
佛教	ဗုဒ္ဓဘာသာ	电视	တယ်လီဗစ်ရှင်း
我	ငါ	市场	ဈေး
神道	ရှင်တိုဘာသာ	通货	ငွေကြေး
基督教	ခရစ်ယာန်ဘာသာ	股票	စတော့ခ်
印度教	ဟိန္ဒူဘာသာ	经济学	စောဂဗေဒ
总理	ဝန်ကြီးချုပ်	生物化学	ဇီဝဓာတု
伤心	ဝမ်းနည်း	病理学	ရောဂါဗေဒ
东盟宣言	အာဆီယံနေ့	天文学	နက္ခတ္တဗေဒ
政治	နိုင်ငံရေး	代数	အက္ခရာသခ်ာ
打嗝	ကြိုးထိုးခြင်း	大学	တက္ကသိုလ်

本文将准确率 $P@N$(选取 N 个候选词时抽取的准确率)作为衡量双语词典好坏的评价指标。其中通过随机抽取验证词典的 3 000 对汉缅双语词汇，PT 为测试集单词的数量，$C(w_i)$ 为抽取方法在单词 w_i 上的抽取结果，若抽取的双语词对正确则取 1，否则取 0，具体计算如式(6)所示。

$$P@N=\frac{\sum_i^{PT}\|C(w_i)\|}{PT} \tag{6}$$

3.2　实验参数设置

在实验中，利用汉缅可比语料通过 Word2Vec 生成汉语、缅语 300 维单语词向量，利用多语言 BERT 模型将汉语、缅语候选词的上下文特征词向量转换为 768 维的词向量，汉缅双语词向量的映射矩阵为 $\boldsymbol{W}$，汉缅的种子词典为 5 000 对，超参数 λ 设定为 0.8。

3.3　实验结果与分析

本文为了体现从可比语料中抽取双语词典方法的效果，设计了四组对比实验。

实验一　不同模型方法对实验结果的影响

为了验证本文方法的有效性，将该方法与不同基线模型进行比较。

1) 基于种子词典的方法；

2) 基于对抗网络的方法；

3) 基于 LDA 的方法；

4) 基于种子词典＋CBW 的方法。

使用同样的数据集进行训练和测试，分别记录每组实验在 $P@1$(即抽取 1 个候选词)时的准确率，实验结果如表 3 所示。

表 3　本文方法与传统方法构建双语词典的准确率

方法	$P@1$/%
基于种子词典的方法	36.93
基于对抗网络的方法	37.30
基于 LDA 的方法	38.24
基于种子词典＋CBW 的方法	44.12
本文方法	**47.69**

分析表 3 的实验数据可知，本文方法可以有效抽取双语词典，效果优于其他几种传统的方法。本文方法优于基于种子词典的方法，主要是因为现有

① https://zh.wikipedia.org/wiki/

的汉缅双语词典规模较小，在小规模种子词典的约束下，难以学习到效果较好的双语映射关系。基于LDA的双语词典抽取方法，主要利用LDA模型抽取到篇章级可比文档中的主题信息，并将这种主题信息视为词语间的语义关系。但是这种方法往往只能抽取到小规模的主题词，而忽视了种子词典的作用。基于种子词典＋CBW的方法效果低于本文方法，主要是因为BERT模型会更加充分地关注上下文的语义信息，可以得到更好的上下文特征表示向量，并将此上下文特征表示向量作为抽取双语词汇的约束特征来提高抽取效果。另外，本文的方法在小规模的种子词典的基础上，通过迭代自学习，不断地扩充种子词典的规模，也能同时学习得到更好的双语映射关系。

实验二　候选词个数对实验结果的影响

为验证方法的准确率与抽取的候选词个数之间的关系，实验还比较了 $P@1$、$P@5$ 和 $P@10$ 的准确率。具体实验结果如表4所示。

表4　不同方法在不同 $P@N$ 值下的准确率

	$P@1$/%	$P@5$/%	$P@10$/%
本文方法	47.69	49.76	**51.24**

分析表4可知，本文方法的准确率均随候选词的增多而逐渐提高，候选词数量仅为1时便可获得较高的准确率；当候选词达到10个时，最高准确率可以达到51.24%，这进一步说明了不同语言在词向量空间中的同构性。在本组实验中我们还观察到一词多义的情况，在句子“太阳，或称日，是太阳系中心的恒星”中，词语“日”在候选词为1和5时，抽取得到意思错误的缅语词“ꩻ(天)”，在候选词为10时，抽取得到意思正确的缅语词“ꩻ(太阳)”。

实验三　不同窗口大小对实验结果的影响

为了验证汉-缅词汇结合上下文窗口的大与本文方法之间的关系，实验还比较了候选词集合个数 N 分别为1、5、10，上下文窗口大小 k 分别为1、2、3、4时的准确率。实验结果如图2所示。

分析图2可知，本文方法的准确率随着上下文窗口大小的增加而逐渐提高，在窗口大小为3时达到最大值。实验证明了上下文信息对双语词向量有着约束作用，能有效提高抽取的准确率，但当上下文窗口大小达到一定程度后，距离目标词太远的词语会对目标词上下文信息进行干扰，产生负面影响。

实验四　不同 λ 值对实验结果的影响

为了研究本文方法的抽取准确率与超参数λ的关系，实验还比较了在候选词集合个数 N 分别为1、5、10时，超参数λ设置为0.6、0.7、0.8、0.9时的准确率。实验结果如表5所示。

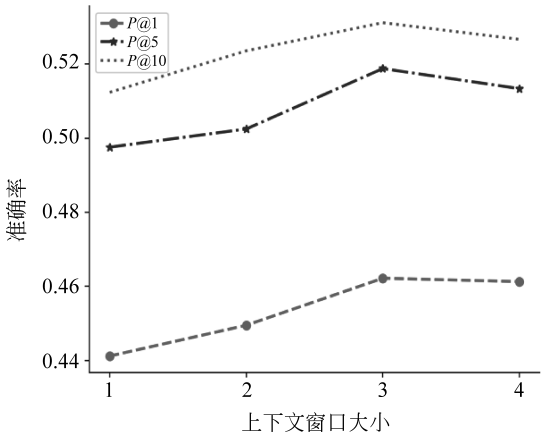

图2　本文方法在不同上下文窗口大小下的准确率

表5　本文方法在不同 λ 值下的准确率

λ	0.6	0.7	0.8	0.9
$P@1$/%	45.96	46.22	47.69	47.48
$P@5$/%	47.01	46.97	49.76	49.14
$P@10$/%	47.69	47.87	**51.24**	50.45

从表5可以看出超参数λ的大小对实验结果有显著的影响。通过上述实验可以得知，超参数λ从0.6增长到0.8时，实验抽取准确率随着其增长也增长，λ越大，上下文信息对平行词抽取的约束效力越高，但当λ增长到0.9时，实验准确率却降低了，证明词语本身的信息约束在抽取中也起着重要的作用。

4　结论

针对汉缅双语语料数据缺乏，当前主流方法无法利用汉缅可比语料中上下文信息的问题，本文提出了一种基于半监督的汉缅双语词典构建方法。首先，汉缅可比语料能有效缓解汉缅双语数据不足导致的模型性能不佳的问题。其次，基于BERT得到抽取词在汉缅文档中的上下文表示，能进一步约束迭代自学习得到的双语词映射表示。本文方法在选取10个缅语候选词时，准确率达到了51.24%。在下一步的研究中，我们将把本文方法应用在更多的东南亚低资源语言中，并将进一步关注汉语-低资源语言对中存在的一词多义的现象，以提高汉语-低资源语言的双语词典质量。

参考文献

[1] 张檬，刘洋，孙茂松. 基于非平行语料的双语词典构建[J]. 中国科学：信息科学，2018，48(05)：84-93.

[2] Artetxe M，Labaka G，Agirre E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics，2018：789-798.

[3] Devlin J，Chang M W，Lee K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Minnesota：Association for Computational Linguistics，2019：4171-4186.

[4] Vulic I，Moens M F. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexiconinduction[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing，Beijing，2015：719-725.

[5] Chandar A P S，Lauly S，Larochelle H，et al. An autoencoder approach to learning bilingual word representations[C]//Proceedings of Advances in Neural Information Processing Systems，2014，27：1853-1861.

[6] Gouws S，Bengio Y，Corrado G. Bilbowa：Fast bilingual distributed representations without word alignments[C]//Proceedings of the 32nd International Conference on Machine Learning，2015：748-756.

[7] Wick M，Kanani P，Pocock A. Minimally-constrained multilingual embeddings via artificial code-switching[C]//Proceedings of the AAAI Conference on Artificial Intelligence，2016：30(1).

[8] Duong L，Kanayama H，Ma T，et al. Learningcrosslingual word embeddings without bilingual corpora[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing，2016：1285-1295.

[9] Cao H，Zhao T，Zhang S，et al. A distribution-based model to learn bilingual word embeddings[C]//Proceedings of the 26th International Conference on Computational Linguistics. Osaka，Japan，2016：1818-1827.

[10] Zhang M，Liu Y，Luan H，et al. Adversarial training for unsupervised bilingual lexicon induction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver，Canada，2017：1959-1970.

[11] Conneau A，Lample G，Ranzato M A，et al. Word translation without parallel data[C]//Proceedings of the International Conference on Learning Representations，2018：74-88.

[12] Artetxe M，Labaka G，Agirre E. Bilingual lexicon induction through unsupervised machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence，Italy，2019：5002-5007.

[13] Riley P，Gildea D. Unsupervised bilingual lexicon induction across writing systems[J]. arXiv preprint arXiv：2002.00037，2020.

[14] Mohiuddin M T，Bari M S，Joty S. LNMap：Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing（EMNLP），2020：2712-2723.

[15] AUNG HLA MOE.基于汉-缅双语语料的双语实体抽取方法研究[D]. 昆明：昆明理工大学硕士学位论文，2018.

[16] Thu Y K，Pa W P，Utiyama M，et al. Introducing the asian language treebank（alt）.[C]//Proceedings of the 10th International Conference on Language Resources and Evaluation，2016：1574-1578.

[17] Mikolov T，Chen K，Corrado G，et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv：1301.3781，2013.

毛存礼(1977—)，博士，副教授，硕士生导师，主要研究领域为自然语言处理、信息检索、机器翻译。

E-mail：maocunli@163.com

王红斌(1984—)，通信作者，博士，副教授，硕士生导师，主要研究领域为自然语言处理。

E-mail：whbin2007@126.com

陆杉(1994—)，硕士研究生，主要研究领域为自然语言处理、机器翻译。

E-mail：188301710@qq.com