文章编号：1003-0077(2022)10-0088-09

融入词集合信息的跨境民族文化实体识别方法

杨振平[1,2]，毛存礼[1,2]，雷雄丽[2,3]，高盛祥[1,2]，陆 杉[1,2]，张勇丙[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；
3. 昆明冶金高等专科学校，云南 昆明 650500)

摘 要：跨境民族文化领域实体通常由描述民族文化特征的领域词汇组合构成，使用当前主流的基于字符表征的实体识别方法会面临领域实体边界模糊问题，造成实体识别错误。为此，该文提出一种融入词集合信息的跨境民族文化实体识别方法，利用领域词典获取的词集合增强领域实体的词边界和词语义信息。首先，构建跨境民族文化领域词典，用于获取词集合信息；其次，通过词集合注意力机制获取词集合向量之间的权重，并融入位置编码增强词集合位置信息；最后，在特征提取层融入词集合信息，增强领域实体边界信息并缓解仅使用字符特征表示所带来的词语义缺失问题。实验结果表明，在跨境民族文化文本数据集上所提出方法相比于基线方法的 F_1 值提升了 2.71%。

关键词：跨境民族文化；实体识别；词集合信息；领域词典；注意力机制

中图分类号：TP391　　**文献标识码**：A

Cross-border National Cultural Entity Recognition Method with Word Set Information

YANG Zhenping[1,2], MAO Cunli[1,2], LEI Xiongli[2,3], GAO Shengxiang[1,2], LU Shan[1,2], ZHANG Yongbing[1,2]

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

Abstract: Cross-border national cultural entities are usually composed of domain words that describe national cultural characteristics. This paper proposes a cross-border national cultural entity recognition method with word set information obtained from domain lexicon. Firstly, a cross-border national cultural domain lexicon is constructed to obtain the word set information. Secondly, the weight between the word set vectors is obtained through attention mechanism, and the positional encoding is adopted. Finally, the word set information is incorporated into the feature extraction layer to enhance the domain entity boundary information and alleviate the problem of word information loss caused by using only character features. Experimental results show that, the F_1 value of the proposed method is improved by 2.71% compared with the baseline method.

Keywords: cross-border national culture; entity recognition; word set information; domain lexicon; attention mechanism

收稿日期：2021-11-01　**定稿日期**：2021-11-23

基金项目：国家自然科学基金(61732005, 61866019, 61761026, 61972186)；云南省应用基础研究计划重点项目(2019FA023)；云南特色产业数字化研究与应用示范(202002AD080001)；云南省中青年学术和技术带头人后备人才项目(2019HB006)

0　引言

跨境民族[1]是指拥有共同的族源关系，后因迁徙或国界变动等原因分别居住在两个或两个国家以上的民族，跨境民族之间语言、习俗、文化等方面基本相同。实体识别是自然语言处理领域的重要任务，需要确定实体边界并将其分类为预定义的实体类型。利用实体识别技术从互联网中自动识别跨境民族文化实体，有助于扩充领域知识图谱和为信息检索提供支撑。因此，研究如何利用深度学习的方法对跨境民族文化文本进行实体识别成为了亟待解决的问题。

目前，序列标注方法在实体识别任务中得到了广泛的应用[2-3]，最新的实体识别模型引入了词汇信息。Zhang 等[4]提出了 Lattice-LSTM 模型，首次将中文词汇信息整合到基于字符的神经网络模型中，利用多粒度信息进行实体识别。Gui 等[5]提出了 LR-CNN 模型，利用句子中平行匹配的所有字符和潜在外部词汇信息进行建模，Ma 等[6]提出 SoftLexicon 方法在字符编码层引入了词信息，所匹配的词来源于上下文句子中词粒度的序列信息，现有的实体识别研究大多集中在通用领域实体识别。由于领域数据集中存在大量的领域实体，在缺少领域知识的辅助下，以上方法不适用于跨境民族文化领域实体识别任务。

跨境民族文化文本中存在大量词边界模糊的领域实体，示例如表 1 所示。例如，饮食文化中的“**毫糯索**”词边界为“**索**”，但是在实体识别中以“**糯**”为词边界，使得现有方法将“**毫糯索**”错误识别为“**毫糯**”。不同词语组合构成的实体存在不同的词边界，例如“**芒勐町佛塔**”由“**芒勐町**”和“**佛塔**”组合构成，在没有领域知识的辅助下，模型无法精准识别出组合实体边界。

表 1　跨境民族文化文本示例

跨境民族文化	文本分析
饮食文化	傣族有许多独具特色的糯米制品：如香竹饭、**毫糯索**、**毫崩**、**千层年糕**等
文艺文化	傣族的民间乐器主要有象脚鼓、**铓锣**、葫芦笙、**铙钹**、箫等。彝族最具代表性的丧事中表演的舞蹈“**孜额且**”
节日文化	西双版纳泼水节的第一天傣语称“**桑刊日**”
建筑文化	**允燕塔**，傣语为**芒勐町佛塔**

续表

跨境民族文化	文本分析
习俗文化	这种习俗在许多佛寺里失传，只有在**曼春满佛寺**，保留了**赕嘎腊洼**这种习俗
位置	**喊沙寨**正在开发乡村旅游，所展示都是傣族居民的日常生活

针对以上问题，本文在 Ma 等[6]所提出的 Soft-Lexicon 方法基础上，提出融入词集合信息的跨境民族文化实体识别方法，本文方法与 Ma 等[6]方法不同的是：为了使模型更好地结合领域知识，本文构建了跨境民族文化领域词典，通过字符与领域词典匹配获取词集合，利用词集合注意力机制和位置编码增强词集合信息，将其融入字符表征中，解决了基于字符表征中存在实体边界模糊和语义缺失问题。

本文的贡献具体如下：

（1）构建了 15.47 万的跨境民族文化领域词典，利用词集合匹配方法从领域词典中获取词集合信息，将其融入到实体识别模型中，缓解跨境民族文化领域实体边界模糊问题。

（2）基于词集合注意力机制获取词集合向量之间的重要程度，利用位置编码增强词集合的位置信息，通过以上方法有效增强词集合信息。

1　相关工作

传统的实体识别方法分为无监督学习方法和基于特征模板的方法。无监督学习方法从语料库上获得词汇资源、词汇模型和统计信息，使用聚类[7]推断命名实体类型。在医学领域中，Wang[8]等利用模板增强的方法并依赖于语料库对生物医学领域进行实体识别。传统的实体识别方法过度依赖人工构建匹配规则并且耗费大量时间筛选有效特征。

随着深度学习方法的兴起，端到端的神经网络模型成为了实体识别的主流方法。Lample 等[9]将 Bi-LSTM 与条件随机场模型（Conditional Random Field，CRF）[10]相结合，弥补了 LSTM 未能考虑样本输出之间关系的缺陷。Hovy 等[11]利用字符级卷积神经网络表示字符特征并构建了字符级 CNN-CRF 模型。李明扬等[12]利用自注意力机制捕获上下文相关信息，更好地理解句子结构。然而这些方法以字符为单位作为输入特征，未结合词信息导致模型很难获取词边界信息和文本深层语义信息。

将词汇信息融入图神经网络的实体识别模型[13]能够打破字符级表示中难以获取词信息的问题，Sui等[14]提出的LGN模型采用协作图网络来直接有效地整合中文词汇知识，解决整合自匹配词汇词和最近的上下文词汇词的问题，使字符和词汇之间有更充分的联系。Li等[15]提出FLAT模型，将栅格结构转换成平滑结构，可以直接建模字符与所有匹配的词汇信息之间交互。

以上方法在通用领域的实体识别上取得了不错的效果，给跨境民族文化实体识别任务提供了较好的思路，但在跨境民族文化实体识别中仍然存在领域实体边界模糊问题以及领域知识如何有效地融入实体识别模型的问题。

2 构建领域词典和词集合匹配方法

2.1 领域词典构建

本文从跨境民族网站和社交媒体上收集了大量的跨境民族文化语料用于抽取领域词语，具体抽取方法如下：

(1) **统计方法** 首先对每一份语料进行预处理，然后统计每一份语料中TF-IDF[16]值靠前的50个词$W=\{W_i|1\leqslant i\leqslant 50\}$，$W_i$表示50个词中第$i$个词，最后计算词$W_i$的归一化词频$tf_i$，当满足$tf_i\leqslant n$时($n$为预先设置的阈值)，将词$W_i$提取到领域词典中。

(2) **人工方法** 利用百度百科收集关于跨境民族文化的词条，通过人工方式对语料中存在边界模糊的专业词语进行提取。

本文一共收集了5 136个关于跨境民族文化的专业词语，词汇的平均长度为4，词语示例如表2所示。

表2 跨境民族文化领域词语示例

跨境民族文化	领域词语
饮食文化	舂牛干巴丝、酸笋煮鱼、菠萝煮饭、树叶炒鸡蛋、叶包蒸猪肉
文艺文化	朗高罕、京省勐晃、傣风神韵、勐巴娜西、片镲、嘎光、娥并与桑洛
节日文化	桑刊日、赶花街、嘎光露、豪洼、叫谷魂
建筑文化	嘎洒寨缅寺、曼飞龙塔、土掌房、高基座佛塔、官家竹楼、干栏式建筑
习俗文化	胆汁刺文、赕文化、踩水碓、赕佛、摆帕拉
位置	芒滚寨、熔岩坝、知青园、姐告、丘陵坝

如表2中领域词语由不同的词组合构成，组合词语在实体识别中的边界模糊，如“嘎洒寨缅寺”由“嘎洒寨”和“缅寺”组合构成，在常规实体识别中就会将其错误地识别为位置和建筑的实体类型，实际上却为建筑类型。将提取的领域词语添加到分词工具中，避免对语料中的词语错误切分，选用开源库gensim① 中的Word2Vec[17]词向量模型训练跨境民族文化语料，获取由15.47万个词向量组成的领域词典，词向量维度为50维。

2.2 词集合定义

词集合是将输入文本中的每个字符通过领域词典匹配得到所有匹配的词所组成的集合，根据字符在所匹配词中的位置不同分为四种词集合类型：字符位于词的头部(B)、字符位于词的内部(M)、字符位于词的尾部(E)和单个字符(S)四个词集合标签。

2.3 词集合匹配方法

如图1所示，“傣族的菠萝紫米饭”中的字符“米”通过式(1)匹配到的词集合为“B=$\{$**米**饭$\}$”“M=$\{$紫**米**饭，菠萝紫**米**饭$\}$”“E=$\{$紫**米**$\}$”和“S=$\{$**米**$\}$”，如果字符在领域词典中没有匹配词，就将该类型的词集合定义为“None”。

跨境民族文化文本$s=\{c_1,c_2,\cdots,c_n\}\in V_c$中的字符$c_i$所匹配的四种词集合由式(1)所示。

$$\begin{cases} B(c_i)=\{w_{i,k},\forall w_{i,k}\in V_w, i<k\leqslant n\} \\ M(c_i)=\{w_{j,k},\forall w_{j,k}\in V_w, 1\leqslant j<i<k\leqslant n\} \\ E(c_i)=\{w_{j,i},\forall w_{j,i}\in V_w, 1\leqslant j<i\} \\ S(c_i)=\{c_i,\exists c_{i,}\in V_w\} \end{cases} \tag{1}$$

其中，V_w表示预先构建好的领域词典，w表示在领域词典中存在的词语，i表示字符的位置，j、k表示字符开始和结束位置。

3 融入词集合信息的跨境民族文化实体识别方法

本文提出了融入词集合信息的跨境民族文化实体识别方法，其模型框架如图1所示，包含了以下六个部分。

(1) **输入层** 该层对输入的文本进行分字表示，作为后续字符与领域词典匹配的数据。

① radimrehurek.com/gensim/models/coherencemodel.html

（2）**字符编码层**　该层对输入模型的跨境民族文化句子序列使用 BERT[18] 语言模型来获得句子中字符向量表示。

（3）**词集合向量编码层**　该层将词集合向量融入到 BERT 编码后的字符向量中。

（4）**双向门控循环单元（Bidirectional Gated Recurrent Unit，BiGRU）[19]编码层**　该层通过 BiGRU 门结构对特征向量进行信息筛选，获取特征向量之间长距离依赖信息。

（5）**自注意力机制层**　该层采用自注意力机制区分 BiGRU 编码输出的特征向量之间的重要程度。

（6）**标签预测层**　该层通过 CRF 预测跨境民族文化文本中每个字符对应的实体标签。

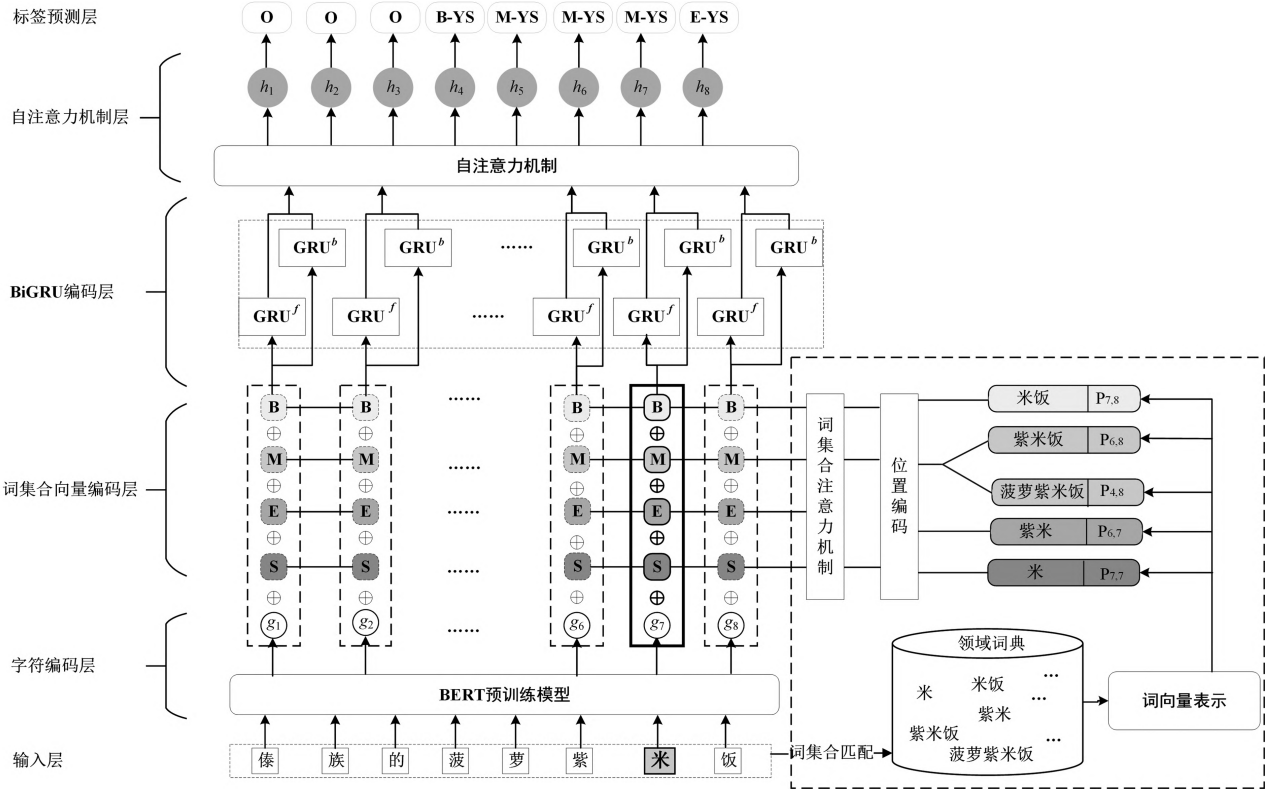

图 1　融入词集合信息的跨境民族文化实体识别模型图

3.1　基于 BERT 的字符向量表示

图 1 在字符编码层采用 BERT 预训练模型对输入的文本进行字符向量表示。文本被视为一个字符序列 $s=\{c_1,c_2,\cdots,c_n\}\in V_c$，其中，$V_c$ 是构建的字典，c_i 是长度为 n 的句子 s 中第 i 个字，c_i 通过 BERT 训练得到字符向量表示，由式(2)所示。

$$\boldsymbol{g}_i=\mathrm{BERT}(c_i) \tag{2}$$

其中，$\boldsymbol{g}_i$ 表示文本中第 i 个字符的字符向量表示，BERT(·)为预训练模型。

3.2　融入词集合信息的跨境民族文化文本表示方法

如图 1 所示，通过字符与词典匹配获取词集合，利用词集合注意力机制获取词集合向量之间的重要程度，通过位置编码增强不同词集合类型之间的区分度，将得到的词集合向量与字符向量进行融合。

3.2.1　词集合向量表示

本文通过构建的领域词典对数据集进行分词后统计每个词的词频。首先对所匹配词的词频加权求和得到四种类型的词频 Z，然后将词频融入词向量中，最后将每种类型中的词向量进行拼接得到词集合向量，由式(3)所示。

$$\begin{cases} Z=\sum\limits_{w\in B\cup M\cup E\cup S} z(w_i) \\ \boldsymbol{v}_i(L)=\dfrac{4}{Z}\sum\limits_{w_i\in L} z(w_i)\boldsymbol{e}(w_i) \end{cases} \tag{3}$$

其中，$z(w_i)$ 为词 w_i 在数据集中统计的词频，$\boldsymbol{e}(w_i)$表示词 w_i 对应的词向量。L 表示$\{B,M,E,S\}$四种类型中的一种，$\boldsymbol{v}_i(L)$为最终得到的词集合向量。

3.2.2 位置编码

文本中的字符位置包含词边界信息，字符的位置不同，所匹配到的词也有所不同，当词集合向量距离非常近时，这些集合向量就不能明显地区分词集合类型，导致后续无法充分利用词集合信息。因此本文在词集合向量中添加了位置编码，如图2所示。

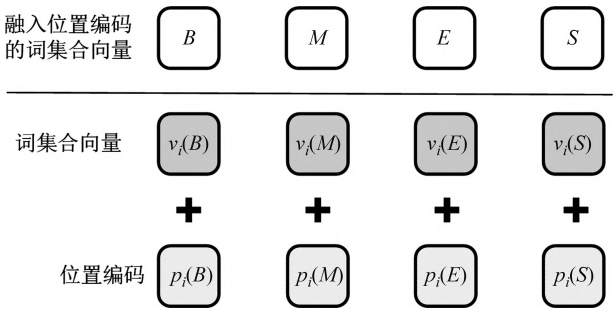

图2 词集合位置嵌入层

本文采用one-hot向量对位置信息向量化表示，添加了训练权重矩阵，以更好地区分词集合向量，位置编码如式(4)所示。

$$\begin{cases}\boldsymbol{v}_i(B)=\boldsymbol{p}_i(B)\boldsymbol{W}^L+\boldsymbol{v}_i(B)\\ \boldsymbol{v}_i(M)=\boldsymbol{p}_i(M)\boldsymbol{W}^L+\boldsymbol{v}_i(M)\\ \boldsymbol{v}_i(E)=\boldsymbol{p}_i(E)\boldsymbol{W}^L+\boldsymbol{v}_i(E)\\ \boldsymbol{v}_i(S)=\boldsymbol{p}_i(S)\boldsymbol{W}^L+\boldsymbol{v}_i(S)\end{cases}\tag{4}$$

其中，$\boldsymbol{W}^L$ 为 $4\times d^w$ 的训练权重矩阵，$\boldsymbol{p}_i(B)=[1,0,0,0]$，$\boldsymbol{p}_i(M)=[0,1,0,0]$，$\boldsymbol{p}_i(E)=[0,0,1,0]$，$\boldsymbol{p}_i(S)=[0,0,0,1]$。

3.2.3 词集合信息融入

得到的四种词集合向量之间重要程度有所不同，如在词集合“$B=\{$米饭$\}$”“$M=\{$紫米饭，菠萝紫米饭$\}$”“$E=\{$紫米$\}$”和“$S=\{$米$\}$”中，M标签的词集合中含有领域实体，因此 M 的重要程度比其它三种词集合更重。为了充分考虑四种词集合向量之间的重要程度，采用词集合注意力机制(图3)计算词集合向量之间的权重。

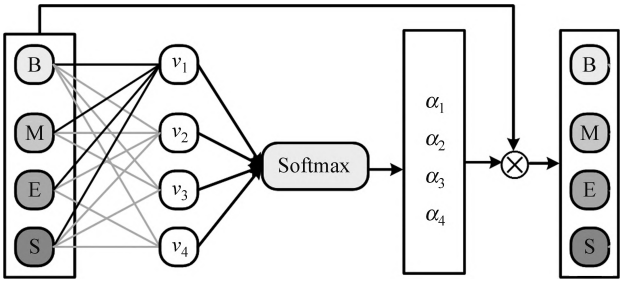

图3 词集合注意力机制

如图3所示，本文通过词集合注意力得到参数矩阵 $\boldsymbol{W}^v$，然后通过归一化函数输出最终的权重向量，如式(5)所示。

$$\begin{aligned}\boldsymbol{V}_i&=\boldsymbol{W}^v\begin{bmatrix}\boldsymbol{v}_i(B)\\ \boldsymbol{v}_i(M)\\ \boldsymbol{v}_i(E)\\ \boldsymbol{v}_i(S)\end{bmatrix}^{\mathrm{T}}+\boldsymbol{b}^v\\ \alpha_i&=\mathrm{softmax}(\boldsymbol{V}_i)\end{aligned}\tag{5}$$

其中，$\boldsymbol{W}^v$ 为参数矩阵，$d^w=50$，$\boldsymbol{b}^v$ 为偏移量，softmax为归一化操作。最后得到维度为 1×4 且取值范围为(0,1)的权重向量 $\boldsymbol{\alpha}_i$。

将每个字符向量与该字符所对应的四种词集合向量组合成最终的向量表示，如式(6)所示。

$$\begin{cases}\boldsymbol{e}_i(B,M,E,S)=\boldsymbol{\alpha}_i\otimes[\boldsymbol{v}_i(B);\boldsymbol{v}_i(M);\boldsymbol{v}_i(E);\boldsymbol{v}_i(S)]\\ \boldsymbol{x}_i=[\boldsymbol{g}_i;\boldsymbol{e}_i(B,M,E,S)]\end{cases}\tag{6}$$

其中，$\boldsymbol{\alpha}_i$ 为式(5)计算的权重向量，$\boldsymbol{e}(B,M,E,S)$表示四种词集合拼接的特征向量，$\boldsymbol{x}_i$ 表示融入词集合信息的特征向量，$\boldsymbol{g}_i$ 为式(2)中的字符向量。

3.3 跨境民族文化文本特征抽取方法

针对文本上下文依赖的问题，本文选用BiGRU对文本进行特征提取。将 $\boldsymbol{x}_i$ 输入到BiGRU中的重置门与更新门中，重置门控制信息丢失，由式(7)计算；更新门决定传递多少信息量到下一次的状态，由式(8)计算。在BiGRU中，新的隐藏状态 $\boldsymbol{h}_i$ 是由上一次的隐藏状态 $\boldsymbol{h}_{i-1}$ 和当前输入 $\boldsymbol{x}_i$ 通过式(9)计算得到。

$$\boldsymbol{r}_i=\sigma(\boldsymbol{W}^r\cdot[\boldsymbol{x}_i,\boldsymbol{h}_{i-1}])\tag{7}$$

$$\boldsymbol{u}_i=\sigma(W^u\cdot[\boldsymbol{x}_i,\boldsymbol{h}_{i-1}])\tag{8}$$

$$\begin{cases}\widetilde{\boldsymbol{h}}_i=\tanh(\boldsymbol{W}^{\widetilde{h}}\cdot[\boldsymbol{r}_i\odot\boldsymbol{h}_{i-1},\boldsymbol{x}_i])\\ \boldsymbol{h}_i=(1-\boldsymbol{u}_i)\odot\boldsymbol{h}_{i-1}+\boldsymbol{u}_i\odot\widetilde{\boldsymbol{h}}_i\end{cases}\tag{9}$$

其中，σ 是sigmoid激活函数，$\boldsymbol{x}_i$ 为融入领域词典信息的特征向量，$\boldsymbol{h}_{i-1}$为上一个时刻的隐藏状态，$\boldsymbol{r}_i$ 是重置门，$\boldsymbol{u}_i$ 是更新门，$\boldsymbol{W}^r$，$\boldsymbol{W}^u$，$\boldsymbol{W}^{\widetilde{h}}$ 是训练参数，$\tanh(\cdot)$为激活函数。

通过自注意力机制得到特征向量之间的关联性和重要程度，特征向量 $\boldsymbol{h}_i$ 通过自注意力机制计算相应的权值得到输出向量，如式(10)所示。

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{h}_i\times\boldsymbol{W}^Q,\quad\boldsymbol{K}=\boldsymbol{h}_i\times\boldsymbol{W}^K,\quad\boldsymbol{V}=\boldsymbol{h}_i\times\boldsymbol{W}^V\\ \mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\dfrac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}\\ \mathrm{head}_i=\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})\end{cases}\tag{10}$$

其中，$\boldsymbol{W}^{Q}$，$\boldsymbol{W}^{K}$，$\boldsymbol{W}^{V}$ 表示权重参数，$d_k=50$ 为输入特征向量的维度，softmax 为归一化操作。

3.4　实体标签预测层

如图 1 所示，通过文本 $s=\{c_1,c_2,\cdots,c_n\}\in V_c$ 对应预测标签序列 $y=\{y_1,y_2,\cdots,y_n\}$ 的概率计算如式(11)所示。

$$\begin{cases}\boldsymbol{P}_i=\boldsymbol{W}_p\mathbf{head}_i+\boldsymbol{b}_p\\ p(y\mid s)=\dfrac{\exp\left(\sum\limits_{i=1}^{n}(\boldsymbol{P}_{i,y_i}+\boldsymbol{T}_{y_{i-1},y_i})\right)}{\sum\limits_{y\in Y_c}\exp\left(\sum\limits_{i=1}^{n}(\boldsymbol{P}_{i,y_i}+\boldsymbol{T}_{y_{i-1},y_i})\right)}\end{cases}\tag{11}$$

其中，$\boldsymbol{W}_p$，$\boldsymbol{b}_p$ 是计算得分矩阵 $\boldsymbol{P}$ 的参数，$\boldsymbol{T}$ 是一个转移矩阵，$\mathbf{head}_i$ 为自注意力机制层的输出向量。

4　实验分析

4.1　跨境民族文化实体标注数据集

跨境民族文化文本中包含有特定领域独特的命名实体，本文将跨境民族文化实体定义为 6 种类型：位置、节日、饮食、习俗、文艺和建筑，标注数据的平均长度为 57 个字符。

由于本文是跨境民族文化领域实体识别任务，在没有公共数据集的情况下，通过民族网站收集了大量的跨境民族文化语料，人工标注了 15 717 条数据。本文将数据集划分为训练集 10 717 条、验证集 3 000 条，测试集 2 000 条，所划分数据集中的实体分布数量如图 4 所示。

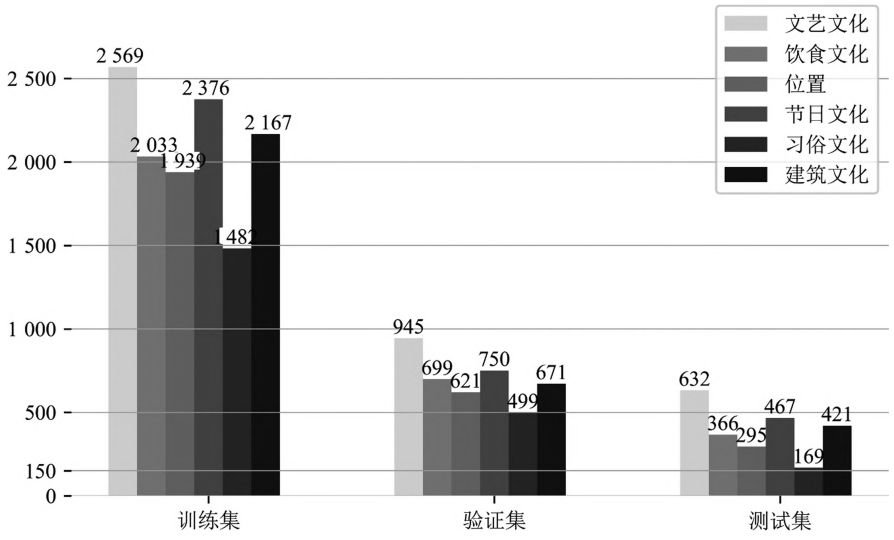

图 4　数据集实体数量统计

4.2　实验参数设置

实验使用 Adamax 优化器来优化所有可训练的参数，实验具体参数设置如表 3 所示。

表 3　模型参数的设置

参数名	参数值
文本最大长度	150
词向量维度	50
隐藏层维度	400
学习率	0.001 5
衰减率	0.05

4.3　实验评价指标

本文的评价指标是通过精确率(Precision)、召回率(Recall)和 F_1 值来对模型进行评估，如式(12)～式(14)所示。

$$\text{Precision}=\frac{A}{A+B}\tag{12}$$

$$\text{Recall}=\frac{A}{A+C}\tag{13}$$

$$F_1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}\tag{14}$$

其中，A 为正确识别实体标签的数量，B 为错误识别实体标签的数量，C 为未被识别到的正确实体标签的数量。

4.4 实验结果与分析

为了验证方法的有效性，本文与现有的实体识别方法做了对比实验，设置了消融实验，验证领域词典、词集合向量编码层以及参数对模型性能的影响，以及验证参数对模型的影响。

实验一：不同方法实验结果对比

与近年来发表的5种方法做了对比实验，FLAT和SoftLexicon(LSTM)所使用的预训练模型为BERT。5种对比方法如下：

(1) Lample等[9]在2016年提出基于字符级的LSTM作为循环神经网络的经典命名实体识别模型；

(2) Zhang等[4]在2018年提出的Lattice-LSTM神经网络的命名实体识别模型；

(3) Gui等[5]在2019年提出的LR-CNN实体识别模型；

(4) Li等[15]在2020年提出的FLAT模型进行实体识别；

(5) Ma等[6]在2020年提出的SoftLexicon(LSTM)方法，本文以该方法作为基线模型。

对比实验结果如表4所示，本文方法在不使用预训练模型BERT情况下与Bi-LSTM+CRF、Lattice-LSTM和LR-CNN模型进行了对比。与Bi-LSTM+CRF相比，本文方法有效地利用词集合信息增强文本上下文语义信息，F_1值提升了6.09%。与Lattice-LSTM和LR-CNN模型相比，本文F_1的方法值分别提升了2.71%、2.01%。因为本文利用自注意力机制对融入词集合信息的特征向量进行选择，有效地获取到重要的特征向量。

表4 对比实验结果 (单位：%)

模型名称	P	R	F_1
Bi-LSTM+CRF[9]	83.59	91.52	87.38
Lattice-LSTM[4]	89.08	92.52	90.76
LR-CNN[5]	92.81	90.15	91.46
BERT+Bi-LSTM+CRF	86.86	92.39	89.54
FLAT+BERT[15]	92.76	**94.45**	93.59
SoftLexicon(LSTM)+BERT[6]	90.68	93.39	92.01
本文方法(-BERT)	94.12	92.82	93.47
本文方法	**95.56**	94.01	**94.72**

本文方法与BERT+Bi-LSTM+CRF和FLAT+BERT模型相比，F_1值分别提升了5.18%和1.13%。本文方法利用词集合信息增强文本上下文语义信息，使实体识别效果得到显著改善。FLAT模型的召回率略高于本文方法，因为FLAT模型通过多层位置编码将外部词汇信息进行编码，其编码结构能够更好地融入Transformer模型中。考虑到BiGRU只有两个门控机制，能够有效减少训练参数，提高模型的训练效率。因此在综合考虑训练效率和实验结果的F_1值情况下，本文选择BiGRU作为特征提取层。

本文方法较基线模型的F_1值提升了2.71%，本文方法增加了词集合注意力机制和位置编码，能够有效地获取四种类型词集合向量重要程度并且增强了词集合向量之间的位置信息。

本文方法在使用预训练模型(BERT)的实验中F_1值比不使用预训练模型(BERT)情况下F_1值提升了1.25%，因为BERT能够更好地表示文本语义性。

实验二：融入领域词典对实验结果的影响

本文设置融入Zhang等[4]构建的通用词典①、融入跨境民族文化领域词典以及未融入词典三个实验证明融入领域词典的有效性。

实验结果如图5所示。可以看出，随着迭代次数的增加，三组实验的F_1值都在明显提升，因为随着迭代次数的增加，模型可以学习到更多的特征信息。

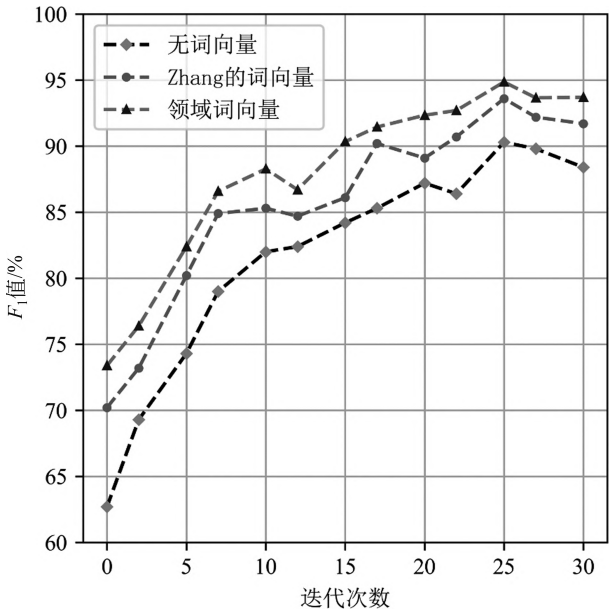

图5 融入领域词典对实验结果的影响

本文对最终迭代次数的F_1值进行了比较，在融入通用词典实验中，F_1值比未融入词典实验提升了3.3%，因为通用词典中包含了词语义信息，本实验证明融入词典对文本语义信息表示的重要性。在

① https://github.com/jiesutd/LatticeLSTM

融入领域词典的实验中 F_1 值比融入 Zhang 的通用词典信息实验提升了 1.2%，因为领域词典中包含跨境民族文化领域词语义信息和实体边界信息。

实验三：位置编码和词集合注意力机制对实验结果的影响

实验三验证位置编码和词集合注意力机制对模型的影响，设计保留位置编码或者词集合注意力机制的消融实验。

如表 5 所示，当模型只保留位置编码时 F_1 值降低了 1.46%，说明词集合注意力对模型效果影响较大，因为基于词集合注意力机制有助于区分四种词集合向量之间的重要程度，进而提升了跨境民族文化实体识别的精度。当模型中只保留词集合注意力机制时，F_1 值降低了 0.74%，说明位置编码可以增强词集合向量之间的位置区分度，模型同时添加位置编码和词集合注意力机制能更充分地获取词集合信息。

表 5　位置编码和词集合注意力对实验的影响

（单位：%）

模型	P	R	F_1
位置编码	94.15	92.39	93.26
词集合注意力机制	94.72	93.25	93.98
位置编码＋词集合注意力机制	**95.56**	**94.01**	**94.72**

实验四：不同的 Dropout_rate 对实验结果的影响

实验四设计了 Dropout_rate 参数分别为 0.3、0.4、0.5、0.6、0.7 的对比实验进行参数的选择。

如表 6 所示，Dropout_rate＝0.3 与 0.4 的精确率都比 Dropout_rate＝0.5 的精确率高，当 Dropout_rate 值较低时模型所学习的参数较多，模型学习能力更强，会造成过拟合现象。当 Dropout_rate＝0.6 与 0.7 时，造成过多的参数丢失导致模型精确率降低，当 Dropout_rate＝0.5 时，模型泛化能力更强。

表 6　不同 Dropout_rate 对实验的影响

（单位：%）

模型	P	R	F_1
Dropout_rate＝0.3	95.70	93.52	94.59
Dropout_rate＝0.4	**96.52**	91.50	93.94
Dropout_rate＝0.5	95.56	**94.01**	**94.72**
Dropout_rate＝0.6	94.78	93.96	94.35
Dropout_rate＝0.7	93.89	92.66	93.27

5　实例分析

为了验证本文方法在跨境民族文化实体识别上的效果，我们在测试数据集中选择了实例进行分析，将 Bi-LSTM＋CRF 方法、SoftLexicon(LSTM)方法和本文方法进行了对比。

如表 7 所示，基于字符级的 Bi-LSTM＋CRF 方法将实体“勐永土锅”错误识别为位置和文艺类型，将“傣家竹编”实体中的“竹编”识别为文艺类型，该方法未融入词集合信息导致模型对边界模糊的词识别效果差。SoftLexicon(LSTM)方法在没有融入领域知识的情况下将“勐永土锅”实体错误识别为位置和文艺类型，本文方法在融入词集合信息后能够准确地识别“勐永土锅”和“傣家竹编”实体。

表 7　实例分析

正确标签	**勐永土锅**(WY)和**傣家竹编**(WY)这些传统制作工艺有着鲜明的民族特色
Bi-LSTM＋CRF	勐永(WZ)土锅(WY)和傣家竹编(WY)这些传统制作工艺有着鲜明的民族特色
SoftLexicon (LSTM)	勐永(WZ)土锅(WY)和**傣家竹编**(WY)这些传统制作工艺有着鲜明的民族特色
本文方法	**勐永土锅**(WY)和**傣家竹编**(WY)这些传统制作工艺有着鲜明的民族特色

通过对预测结果分析，如在句子“掸族风味饮食较为有名的是阔伯，又称红糯米糕。”中，“阔伯”的实体类型未被正确识别，因为该实体未出现在词集合中。

6　结论

针对跨境民族文化实体结构特征，本文提出了融入词集合信息的跨境民族文化实体识别方法，构建了 15.47 万词的领域词典，涵盖了文艺、饮食、节日、建筑以及习俗文化等词。在词集合向量特征提取层中，基于词集合注意力机制方法，增强四种词集合向量之间的重要程度，利用位置编码提升词集合位置信息，使得领域词典更好地融入到字符向量表示中。实验表明，本文方法在跨境民族文化实体识别上有一定提升。下一步将研究跨境民族文化之间存在实体链接的关系，将抽取更多的领域词语，构建更大的专业词汇表，解决未出现在领域词典的词语导致实体识别效果差的问题。

参考文献

[1]　郭家骥.云南周边跨境民族文化交流互动与边疆繁荣稳定[J].云南社会科学，2015,208(06)：122-127.

[2]　Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF for

named entity recognition[C]//Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, Hongkong, 2018: 531-540.
[3] Rei M. Semi-supervised multitask learning for sequence labeling[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017: 2121-2130.
[4] Zhang Y, Yang J. Chinese NER using lattice LSTM [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne,2018: 1554-1564.
[5] Gui T, Ma R T, Zhang Q, et al. CNN-based Chinese NER with lexicon rethinking[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao,2019: 4982-4988.
[6] Ma R T, Peng M L, Zhang Q, et al. Simplify the usage of lexicon in Chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle,2020: 5951-5960.
[7] Nadeau D, Sekine S. A survey of named entity recognition and classification [J]. Lingvisticae Investigationes,2007, 30(1): 3-26.
[8] Wang X, Zhang Y, Li Q, et al. PENNER: Pattern-enhanced nested named entity recognition in biomedical literature[C]//Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. Madrid, Spain,2018: 540-547.
[9] Lample G. Neural architectures for named entity recognition[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, California, 2016: 260-270.
[10] Lafferty J, Mccallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the 18th International Conference on Machine Learning,2001: 282-289.
[11] Hovy E, Ma X Z. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, 2016: 1064-1074.
[12] 李明扬,孔芳. 融入自注意力机制的社交媒体命名实体识别[J].清华大学学报(自然科学版),2019, 59(6): 461-467.
[13] Gui T, Zou Y C, Zhang Q, et al. A lexicon-based graph neural network for Chinese NER[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hongkong, 2019: 1039-1049.
[14] Sui D B, Chen Y B, Liu K, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hongkong, 2019: 3821-3831.
[15] Li X N, Yan H, Qiu X P, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle,2020: 6836-6842.
[16] Robertson S. Understanding inverse document frequency: on the oretical arguments for IDF[J]. Journal of Documentation, 2004,60(5): 503-520.
[17] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J/OL]. arXiv preprint arXiv: 1301.3781, 2013.
[18] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minnesota, 2019: 4171-4186.
[19] Cho K,Merrienboer B V, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014: 1724-1734.

杨振平(1997—),硕士研究生,主要研究领域为自然语言处理、信息抽取。
E-mail: 2412302827@qq.com

雷雄丽(1978—),硕士,讲师,主要研究领域为民族心理学、自然语言处理。
E-mail: 735361482@qq.com

毛存礼(1977—),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、信息检索、机器翻译。
E-mail: maocunli@163.com