山西大学学报(自然科学版)45(4):884－893,2022
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10. 13451/j. sxu. ns. 2021123

融合领域知识图谱的跨境民族文本分类方法

陈春吉[1,2],毛存礼[1,2],雷雄丽[3*],满志博[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 建筑与艺术学院,云南 昆明 650000)

摘　要:跨境民族文本分类任务是跨境民族文化分析中的基础性工作,其目的是将跨境民族文化文本进行归类处理。针对跨境民族文化数据分类面临类别交叉的问题,提出融合领域知识图谱的跨境民族文本分类方法,利用跨境民族文化知识图谱对文本中的跨境民族实体进行语义扩展,通过实体在知识图谱中的类别特征来增强文本的类别语义特征。此外,通过掩码自注意力机制分别对文本的词级、句子级进行特征提取以此得到文本中句子的局部特征和全局特征。实验表明,本文方法在跨境民族文化数据集中相比基线模型的$F1$值提升了11. 9%。

关键词:跨境民族文化;文本分类;领域知识图谱;实体语义扩展
中图分类号:TP391　　**文献标志码:**A　　**文章编号:**0253-2395(2022)04-0884-10

Cross-border Ethnic Text Classification Method Integrating Domain Knowledge Map

CHEN Chunji[1,2], MAO Cunli[1,2], LEI Xiongli[3*], MAN Zhibo[1,2], LU Shan[1,2], ZHANG Yongbing[1,2]
(1. *School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500*, China;*
2. *Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming* 650500*, China;*
3. *School of Architecture and Art, Kunming Metallurgical College, Kunming* 650000*,China*)

Abstract: The task of cross-border ethnic text classification is the basic work in cross-border ethnic cultural analysis, and its purpose is to classify cross-border ethnic cultural texts. Aiming at the problem of cross-category in cross-border ethnic cultural data classification, this paper proposes a cross-border ethnic text classification method based on domain knowledge map, which uses cross-border ethnic cultural knowledge map to expand the semantics of cross-border ethnic entities in the text, and enhances the category semantic features of the text through the category features of entities in the knowledge map. In addition, by using the mask self-attention mechanism, the local features and global features of sentences in the text are obtained by extracting features at word level and sentence level, respectively. The experiments show that the $F1$ value of this method in cross-border ethnic culture data set is improved by 11. 9% compared with the baseline model.

Key words: cross-border ethnic culture; text classification; domain knowledge map; entity semantic extension

收稿日期:2021-09-30;**接受日期:**2021-11-20
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);云南省中青年学术和技术带头人后备人才项目(2019HB006);云南特色产业数字化研究与应用示范(202002AD080001)
作者简介:陈春吉(1995-),硕士研究生,研究方向为自然语言处理。E-mail:1789747432@qq. com
* **通信作者:**雷雄丽(LEI Xiongli),E-mail: 735361482@qq. com
引文格式:陈春吉,毛存礼,雷雄丽,等. 融合领域知识图谱的跨境民族文本分类方法[J]. 山西大学学报(自然科学版),2022,45(4):884-893. DOI:10. 13451/j. sxu. ns. 2021123

0 引言

跨境民族是指居住在不同国家和地区却保留某些原始民族特色,拥有同一民族共同感的民族[1]。例如,中国的傣族和泰国的泰族,在语言、服饰上都有不同程度的改变,却仍保留着原始民族过泼水节,信仰佛教,住竹楼等风俗习惯。跨境民族文化文本是由跨境民族文化中的宗教、服饰、饮食、艺术、建筑、习俗等文化因素组成的文本内容,跨境民族文本分类有助于跨境民族文化的研究与分析。因此,如何利用文本分类技术对跨境民族文化文本数据进行精准分类成为关键。

文本通常包含与该文本类别相关度高的词或句,能够将其发现并提取将极大地提升文本分类的性能。跨境民族文化文本中存在文化类别相互交叉、语义复杂的现象。如表1中的前两条数据,描述泼水节的正文中出现了表示艺术文化的**“象脚鼓舞”**,表示宗教文化的**“浴佛”**,这些活动的出现导致类别的交叉,给分类任务造成一定程度的干扰。此外,表示节日活动的泼水节有**“浴佛节”“宋干节”**以及傣语名称**“楞贺桑勘”**等多种表述方式,使得跨境民族文本数据中存在实体语义相对复杂的问题。

文本内容由正文和标题构成,标题是正文内容的凝练表达,具有提示和补充正文核心内容的作用[2],就跨境民族文本数据而言,部分标题存在跨境民族类别特征信息不明显的现象。此外,跨境民族文化正文数据中存在文本内容较长,特征词众多,跨境民族习俗文化相同的问题,如中国的傣族和泰国的泰族都共同拥有节日文化——泼水节,仅从文字的描述往往出现类别的混淆。

针对以上问题,本文提出了融合领域知识图谱的跨境民族文本分类方法,在Yang等[3]提出的分层注意力文本分类方法上进行了改进,并借鉴Shen等[4]、Bordes等[5]的思想将外部知识信息与文本有效地结合起来辅助正文分类。本文主要贡献如下:

(1)利用跨境民族文化知识图谱对文本中的跨境民族实体进行语义扩展,通过实体在知识图谱中的类别特征来增强文本的类别语义特征。

(2)有效利用标题辅助正文锁定关键词、补充和概括正文的优势将其与正文进行联合,并把提取到的不同层次的特征信息结合到一起辅助分类,初步解决了跨境民族文化类别交叉的问题。

1 相关工作

目前面向文本分类任务主要有两类方法:基于传统机器学习的方法和基于神经网络的深度学习方法。

基于机器学习的文本分类方法,如朴素贝叶斯(Naive Bayes)[6]、支持向量机(Support Vector Machine)[7]以及K近邻(K-Nearest Neighbor)[8]等方法。传统的机器学习文本分类方法通常采用不同类型的机器学习算法作为算法分类器,并结合特征工程进行分类,然而存在难以捕获跨境民族文本深层含义和依赖人工提取跨境民族特征等问题。

基于神经网络的深度学习文本分类方法是当今的主流方法。主要分为基于卷积神经网络(Convolutional Neural Networks)、基于循环神经网络(Recurrent Neural Network)、基于注意力机制(Attention Mechanism)的方法,三者被广泛应

表1 跨境民族文化文本数据样例

Table 1 A sample of cross-border ethnic cultural text data

标题	正文	类别	问题
魅力西双版纳泼水节	节日期间,举行丢包、划龙舟、放高升、放孔明灯、跳象脚鼓舞、跳孔雀舞、斗鸡等活动。泼水的这天早晨,村村寨寨的傣族男女老少穿上节日盛装,提着清水,先到佛寺浴佛……	傣族节日	正文文化概念相互交叉
清迈-气氛最浓的泼水节	在清迈,塔佩门是泼水节(也称为宋干节)指定的狂欢地。本地人和游客都会前往塔佩门享受泼水之乐,这里也是泼水节主战场,现场有舞台、音乐以及户外派对等……	泰族节日	正文文化概念相互交叉
傈僳族民族信仰	……苗人、岱人、热依人在每年耕种完毕,或遇到虫害天灾等不顺心的事,都要供祭傈僳神,祈求保佑庄稼……	傈僳族宗教文化	标题中辅助分类的关键词信息较弱

用在文本分类任务中。(1)基于卷积神经网络的文本分类方法:卷积神经网络由于其提取局部特征的优势被成功引入分类任务中,并取得不错的进展[9-11]。然而,卷积神经网络存在不能捕获长距离依赖信息的缺陷。(2)基于循环神经网络的文本分类方法:循环神经网络因其能处理任意长度数据的特点得到了广泛应用。其中,长短期记忆网络(Long Short-Term Memory)、门控循环单元(Gated Recurrent Unit)等方法都是特殊的循环神经网络,在文本的上下文表示以及特征提取方面各有优势,但存在关键词定位不准确的问题。(3)基于注意力机制的文本分类方法:注意力机制通过捕获数据中的重要信息从而提高文本分类的性能[12-15]。Yang等[3]提出了分层注意力网络(HAN),对单词和句子分别使用注意力机制,分别捕获词级、句子级的重要信息。Wang等[13]提出了可应用于视频动作识别的分层注意力网络。针对跨境民族文本分类任务中存在的问题,考虑采用分层网络强大的语义表征能力表示正文内容,对标题仅采用针对单词级别的注意力机制,以提取标题的重要特征。

综上所述,针对特定领域跨境民族文本分类中存在的语义环境复杂,类别交叉的问题,采用传统的机器学习分类算法或者深度学习的神经网络分类算法容易导致跨境民族文化类别的划分不准确。因此,针对领域中存在的类别交叉问题,可采用知识图谱对文本进行实体语义的扩展从而提高文本的分类效果。

2 跨境民族文化知识图谱构建

跨境民族文化知识图谱通过知识抽取和人工构建的方式构建,其中包含1 064个知识三元组。首先,通过爬虫技术获取维基百科、百度百科以及各大民族网站的跨境民族文化文本数据;然后,分别利用命名实体识别技术和关系抽取技术得到跨境民族文本中的实体以及关系信息,根据(实体,关系,实体)或者(实体,属性,实体)的方式构建三元组并采用人工的方式进行校对,最后,将其通过Neo4j图数据库进行存储。跨境民族文化知识图谱的构建主要包括三个方面:跨境民族文化知识图谱类别体系构建、跨境民族文化实体属性定义、跨境民族文化关系属性定义。跨境民族文化知识图谱存储在Neo4j图数据库中,具体如图1所示。

2.1 跨境民族文化知识图谱类别体系构建

跨境民族文化类别的划分是跨境民族文化知识图谱构建的基础性工作。通过对跨境民族文本数据的具体分析,将跨境民族文化文本数据分为六个大类,具体如表2所示。

由表2可知,每个类别下具体包含一定数量的精细类别。根据表2对各个跨境民族文本

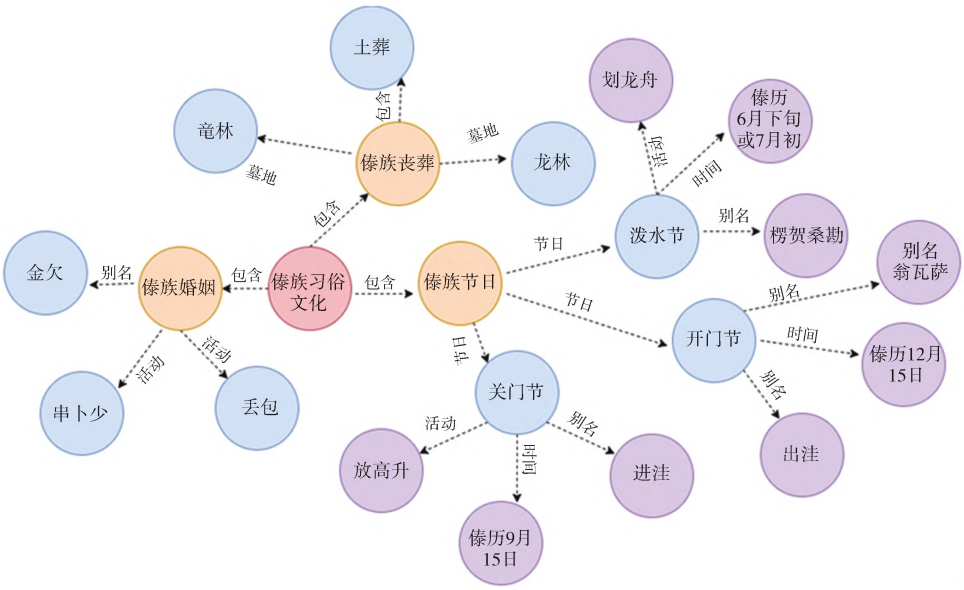

图1 跨境民族知识图谱示例

Fig. 1 Examples of cross-border ethnic knowledge maps

表2 跨境民族文化类别划分

Table 2 Classification of cross-border ethnic cultures

类别名称	类别名称
宗教文化	原始宗教文化、佛教文化
饮食文化	饮品文化、小吃文化
建筑文化	民居建筑文化、寺庙建筑文化
服饰文化	女性服饰文化、男性服饰文化
习俗文化	节日文化、丧葬文化、婚姻文化
艺术文化	音乐文化、舞蹈文化、乐器文化、手工艺术文化

数据的划分可获得某个跨境民族的文化架构,如傣族文化习俗文本数据可以将其分为:傣族节日文化、傣族丧葬文化、傣族婚姻文化等。

2.2 跨境民族文化实体属性定义

跨境民族实体是对跨境民族文化信息的客观表达,跨境民族实体属性的定义是对跨境民族文化领域实体的规范化统一。实体属性的加入使得跨境民族文化知识图谱更加丰富,从而更好地体现跨境民族文化间的差异,具体如表3所示。

表3 跨境民族实体属性定义

Table 3 Definition of cross-border ethnic entity attributes

属性	存储形式	描述
跨境民族实体名称	String	专业名称
跨境民族实体别称	String	除专业名称之外的名称
跨境民族实体描述内容	String	实体的简要描述
跨境民族实体类别标签	String	实体的类别特征

由表3可知,实体的属性主要包括:跨境民族实体名称、跨境民族实体别称、跨境民族实体描述内容以及跨境民族实体类别标签。例如有实体:**“糯米酒”**,则实体的别称为:**“劳毫糯”**,实体的描述内容为**“傣族酒,傣族特制饮品”**,实体类别标签为**“傣族饮食文化,傣族饮品文化”**。

2.3 跨境民族文化关系属性定义

跨境民族关系是跨境民族文化实体间联系的桥梁,是对跨境民族文化领域知识图谱中跨境民族知识的关联整合。跨境民族文化中的实体关系定义主要分为以下几种:包含关系、属性关系、位置关系。跨境民族实体关系的建立使得跨境民族文化知识图谱可视化性能、查询性能得到有效加强。具体如表4所示。

由表4可知,对于跨境民族文化实体**“泼水节”**不仅属于傣族节日,同时拥有**“划龙舟”**这一节日活动,同一实体在不同的语境中所涉及的关系不尽相同,对实体间关系的准确划分将对跨境民族文化知识图谱的构建提供有效保障。

表4 跨境民族实体关系举例

Table 4 Examples of cross-border ethnic entity relationships

实体1	关系	实体2	关系类别
傣族	跨境	泰族	位置关系
傣族节日	包含	泼水节	包含关系
泼水节	活动	划龙舟	属性关系

3 融合领域知识图谱的跨境民族文本分类模型

融合领域知识图谱的跨境民族文本分类模型具体如图2所示,该模型主要由输入层、特征提取层、输出层组成。

3.1 输入层

输入层用来捕获正文和标题中词级语义关系以及每个词对句子的重要性程度。由于正文和标题的结构不同,标题是比较凝练的单句,正文往往由多个句子构成,实体语义相对复杂,为此在正文中融入跨境民族文化知识图谱的实体语义信息辅助正文进行特征编码。本文采用两种不同的方式对标题和正文进行编码,具体如图2中输入层。

(1) 标题词向量编码层

跨境民族文化数据中存在大量的专业名词,普通的分词工具缺少跨境民族文化词库,导致分词效果差,进而影响后续分类工作。例如:“**傣族信奉南传上部座佛教**”,在不使用跨境民族文化词库时得到的结果是“**傣族 信奉 南传 上部座 佛教**”然而“**南传上部座佛教**”是专业名词,不可将其分开,类似的词语在跨境民族文化中还有很多。为此,本文利用人工构建的跨境民族文化词库来辅助jieba工具分词,使用Word2Vec模型对现有的跨境民族文化数据训练得到词向量,并利用跨境民族词向量获得标题中每个词 $w_i, i\in[1,N]$ 的嵌入表征 $x_i\in R^{100}$,其中,N代表标题中词的个数。

采用双向门控循环单元(BiGRU)获得标题的词级表示。BiGRU包含从前向GRU获得标题 x_1 到 x_N 的隐藏向量表示和后向GRU获得 x_N 到 x_1 的隐藏向量表示。具体操作如下:

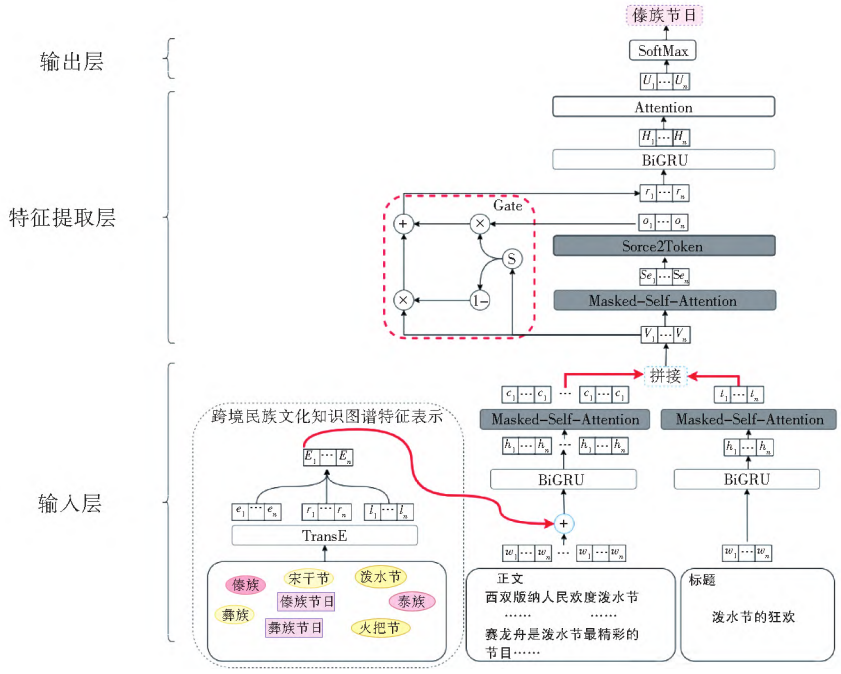

图2 融合领域知识图谱的跨境民族文本分类模型图

Fig. 2 Cross-border ethnic text classification model diagram fused with domain knowledge map

$$h_i = BiGRU(x_i), i \in [1, N], \quad (1)$$

其中 h_i 表示标题中第 i 个单词的前向和后向隐状态信息的结合 $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$。

(2) 标题特征提取层

标题的特征提取采用 Shen 等[4]提出的掩码自注意力机制，与普通的注意力不同的是，掩码自注意力网络是一个多维度的自注意力网络，针对跨境民族文化文本数据中那些在不同语境下意义不同的单词，如："**泼水节**"既可以表示为傣族节日也可表示为泰族节日，通过掩码自注意力机制可以将上下文信息结合起来，从而更好地辅助分类。

首先，为标题中的每个词计算对齐分数，接着采用 Softmax 函数进行归一化计算概率分布，值越大说明标题中的某个词贡献了重要的信息。具体如下所示：

$$f(h_i, h_j)^{bw} = c \cdot \tanh([W^{(1)}h_i + W^{(2)}h_j + b]/c) + M_{ij}^{bw}\mathbf{1}, \quad (2)$$

$$p(z_k = i|h, h_j)^{bw} = \frac{\exp(f(h_i, h_j)^{bw})}{\sum_{i=1}^{n}\exp(f(h_i, h_j)^{bw})}, \quad (3)$$

其中 $f(h_i, h_j) \in R^{d_e}$ 是与输入 h 维度相同的向量，$W^{(1)}, W^{(2)} \in R^{d_e \times d_e}$ 表示权重矩阵；$\tanh(\cdot)$ 表示激活函数；c 表示为标量，实验中通常设置 $c=5$，用来减少参数的数量。$M_{ij}\mathbf{1}$ 中 $\mathbf{1}$ 代表全是 1 的向量，M_{ij} 代表编码时序信息的掩码矩阵。为了获得双向的注意力分数，分别在公式(2)中采用前向的掩码矩阵 $M_{ij}^{f_w}$ 和反向的掩码矩阵 $M_{ij}^{b_w}$，掩码矩阵的具体表示如下：

$$M_{ij}^{f_w} = \begin{cases} 0, i < j \\ -\infty, \text{其他} \end{cases}, \quad (4)$$

$$M_{ij}^{b_w} = \begin{cases} 0, i > j \\ -\infty, \text{其他} \end{cases}。 \quad (5)$$

其次，该注意力机制的输出表示标题中所有词嵌入的加权和，其中权重由 $p_{k_i} \triangleq p(z_k = i|h, h_j)$ 给出，可以将输出写为根据词的重要性采样的令牌期望，即：

$$s^{bw} = [\sum_{i=1}^{n} p_{k_i}{}^{bw} h_{k_i}]_{k=1}^{d_e} = [E_{i \sim p(z_k|h, h_j)^{bw}}(h_{k_i})]_{k=1}^{d_e}, \quad (6)$$

其中 $s^{bw} \in R^{d_e \times N}$ 表示标题中第 j 个序列 h_j 的输出。为了简便，以下的公式中都忽略了下标 k，公式(6)可写为 $s^{bw} = \sum_{i=1}^{n} p^j{}_i \cdot h_i$。

最后，标题的输出 $t \in R^{2d_e \times N}$(前向输出 t^{fw} 和后向输出 t^{bw} 的计算方式一致)由注意力机制的输出 s^{bw} 以及标题的输入 h 通过融合门控机制得到，这将为标题中的每个元素生成一个上下文感知以及时序编码的向量表示。具体如下：

$$F^{bw} = sig \bmod(W^{(f1)}s^{bw} + W^{(f2)}h + b^f), \quad (7)$$

$$t^{hw}=F^{hw}\cdot h+(1-F^{hw})s^{hw}, \tag{8}$$

$$t=[t^{hw}\|t^{fw}]\in R^{2de}, \tag{9}$$

其中$\boldsymbol{W}^{(f1)}$、$\boldsymbol{W}^{(f2)}\in R^{d_e\times d_e}$,$b^f\in R^{d_e}$是融合门控机制中的可学习参数,"‖"表示连接操作,$T=[t_1,t_2,t_3,\cdots,t_n]$。

(3)跨境民族文化知识图谱特征表示

结合Bordes等提出的TransE模型进行跨境民族实体语义的扩展,利用TransE模型将跨境民族文化知识图谱中三元组的实体和关系表示在同一个向量空间中。例如,有跨境民族文化知识三元组(**傣族,傣族节日,关门节**)。首先,分别对头实体、尾实体、关系进行标记处理。将实体标记为(**傣族**,0)、(**关门节**,1);关系标记为(**傣族节日**,0)。实体的标签由实体的别称和实体的类别标签构成,故实体"**傣族**"的标签记为(**掸族**,0_0)和(**跨境民族**,0_1);实体"**关门节**"的标签记为(**进注**,1_0)和(**傣族节日文化**,1_1)。之后,将实体向量、关系向量以及训练数据随机初始化后输入到TransE模型中进行训练。通过不断调整训练参数使其满足公式(10)的目标函数,最终得到实体的向量表示$[e_{傣族},e_{关门节}]$,关系向量表示$[r_{傣族节日}]$以及标签向量(包含别称和类别标签)为$[l_{掸族},l_{跨境民族},l_{进注},l_{傣族节日文化}]$,再把相应的实体向量和关系向量进行对位相加得到实体语义向量为:

$$\boldsymbol{E}_{傣族}=e_{关水节}+r_{傣族节日}+l_{掸族}+l_{跨境民族},$$

$$\boldsymbol{E}_{关门节}=e_{傣族}+r_{傣族节日}+l_{进注}+l_{傣族节日文化}。$$

最后将所有实体的实体语义向量进行存储得到跨境民族实体语义向量表。

在模型训练过程中,为了验证跨境民族三元组的知识表示是否正确,需要根据正确的跨境民族文化知识三元组S构造一些错误的跨境民族文化知识三元组S',并通过公式(10)对其进行衡量,正确的三元组获得较高的分数,错误的三元组获得较低的分数,TransE定义的损失函数具体如下:

$$L=\sum_{(h,r,t)\in S(h',r,t')\in S'(h,r,t)}\sum[\gamma+((h+l_h)+\gamma-(t+l_t))^2-((h'+l_{h'})+\gamma-(t'+l_{t'}))^2]_+, \tag{10}$$

其中S表示正确的跨境民族文化知识三元组;S'表示随机构造的负样例,构造方法为随机将正确的跨境民族文化知识三元组中的实体和关系替换为其他的元素;$l_{h'}$和$l_{t'}$表示随机构造的头实体和尾实体标签;γ是大于0的间隔距离参数;$[\cdot]_+$表示正值函数,当$[\cdot]>0$时,$[\cdot]_+=[\cdot]$,当$[\cdot]<0$时$[\cdot]_+=0$。

(4)跨境民族文本特征编码层

正文的特征编码和标题基本一致,不同的是正文由多个句子构成,而标题是简短的句子。首先,将正文分为多个句子$J_j,j\in[1,M]$,其中M代表句子的个数。采用分词处理得到词向量$\boldsymbol{w}_{jk},k\in[1,N]$,其中$N$代表句子的长度。通过实体在文本中的位置可以将这两种向量对位相加得到文本的词向量,融合过程如下所示:

$$\boldsymbol{A}_i=\boldsymbol{w}_{jk}+\boldsymbol{E}_i, \tag{11}$$

其中$\boldsymbol{E}_i$的维度和$\boldsymbol{w}_{jk}$一致,$\boldsymbol{E}_i$表示通过TransE表征过的实体语义特征信息。例如文本:**泼水节是傣族最隆重的节日**。首先提取两个字以上的词语并通过跨境民族实体语义向量表查找相应的实体语义向量可得到$[\boldsymbol{E}_{傣族},\boldsymbol{E}_{泼水节}]$。之后根据跨境民族词向量得到正文的向量表示$[\boldsymbol{w}_{泼水节},\boldsymbol{w}_{是},\boldsymbol{w}_{傣族},\boldsymbol{w}_{最},\boldsymbol{w}_{隆重},\boldsymbol{w}_{的},\boldsymbol{w}_{节日}]$,最后将文本向量与能查找得到的实体语义向量进行融合可得到

$$[\boldsymbol{w}_{泼水节}+\boldsymbol{E}_{泼水节},\boldsymbol{w}_{是},\boldsymbol{w}_{傣族}+\boldsymbol{E}_{傣族},\boldsymbol{w}_{最},\boldsymbol{w}_{隆重},\boldsymbol{w}_{的},\boldsymbol{w}_{节日}]。$$

使用BiGRU进行编码得到正文隐向量$\boldsymbol{h}_i$,采用掩码自注意力网络(Masked-Self-Attention)进行正文句子特征提取,得到正文词级特征向量表示为:

$$\boldsymbol{C}=[c_1,c_2,c_3,\cdots,c_l],j\in[1,M],$$

之后将正文和标题的特征向量进行融合,

$$\boldsymbol{V}=[\boldsymbol{T},\boldsymbol{C}] \tag{12}$$

其中$\boldsymbol{V}\in R^{(N+M)\times 2d_e}$为融合后的特征向量,$T=[t_1,t_2,t_3,\cdots,t_n]\in R^{N\times 2d_e}$表示标题特征向量,$\boldsymbol{C}\in R^{M\times 2d_e}$表示正文特征向量。

3.2 特征提取层

跨境民族文化文本句子中常存在时间顺序、因果关系和其他的逻辑关系。不同的句子可能包含不同的跨境民族特征信息。因此,特征提取层用来捕获正文和标题之间的上下文依赖关系。具体为图1中的第二部分。

(1)融合标题和正文的跨境民族特征提取层

融合标题和正文的特征提取层负责连接和融合标题和正文中的信息并进行特征的提取。通过3.1节输入层可以得到标题和正文的词级

特征编码矩阵 $\boldsymbol{V}$。将融合了标题和正文的文本特征矩阵输入到掩码自注意力模型获得前向的特征矩阵 $\overrightarrow{\boldsymbol{se}_i}$ 和后向的特征矩阵 $\overleftarrow{\boldsymbol{se}_i}$，通过融合得到特征矩阵 $\boldsymbol{Se}_i=\overrightarrow{\boldsymbol{se}_i}\|\overleftarrow{\boldsymbol{se}_i}$，其中，“‖”表示连接操作。之后将 $\boldsymbol{Se}_i$ 作为输入采用 Sorce2Token 网络进行句子间特征的提取，具体如下所示：

$$\boldsymbol{f}(x_i)=\boldsymbol{W}^T\sigma(\boldsymbol{W}^1\boldsymbol{Se}_i+b^1)+b \tag{13}$$

其中 $\boldsymbol{W}^T,\boldsymbol{W}^1,b^1,b\in R^{4d_e\times 4d_e}$，$\sigma(\cdot)$ 表示激活函数。接着，采用 Softmax 函数进行归一化确定权重，概率矩阵被定义为 $\boldsymbol{p}_{\mathrm{ki}}\triangleq p(z_k=i|\boldsymbol{Se})$，输出为：$o=\sum_{i=1}^{n}p_{.i}\bullet\boldsymbol{Se}_i$。最后，利用门控机制将标题和正文的词级和句子级信息进行联合，具体如下所示：

$$F=sig\,\mathrm{mod}(\boldsymbol{W}^{(f1)}o+\boldsymbol{W}^{(f2)}\boldsymbol{V}+b^f), \tag{14}$$

$$r=F\bullet\boldsymbol{V}+(1-F)o, \tag{15}$$

其中 $\boldsymbol{W}^{(f1)}\in R^{4d_e\times 4d_e}$，$\boldsymbol{W}^{(f2)}\in R^{4d_e\times 4d_e}$，$b^f\in R^{4d_e}$，$o$ 表示融合标题和正文的句子级特征信息，$\boldsymbol{V}$ 表示标题和正文的词级特征信息。

(2) 跨境民族上下文特征融合层

上下文特征融合层通过采用 BiGRU 获得全局信息的上下文编码矩阵。与 3.1 节中词向量特征编码不同的是前者将标题和正文分别进行特征的提取，后者对标题和正文信息联合后的全局特征提取，且对标题和正文之间的交互更加关注，具体如下所示：

$$\boldsymbol{H}_i=BiGRU(r_i), \tag{16}$$

$$u_i=\tanh(\boldsymbol{W}_u\boldsymbol{H}_i+b_w), \tag{17}$$

$$a_i=\frac{\exp(u_i^T,u_w)}{\sum_{i=1}^{n}\exp(u_i^T,u_w)}, \tag{18}$$

$$\boldsymbol{U}_i=a_i\boldsymbol{H}_i, \tag{19}$$

其中 $\boldsymbol{W}_u\in R^{4d_e\times 4d_e}$，$b_w\in R^{4d_e}$。由上述公式可知，首先通过双向循环单元获得隐藏向量表示 $\boldsymbol{H}_i$，之后通过多层感知机计算注意力分数，之后采用 Softmax 进行归一化确定权重 a_i，最终得到文档集的特征编码向量 $\boldsymbol{U}_i$。

3.3 输出层

本层是为了判断跨境民族文化的 28 个类别，通过 Softmax 函数计算属于某个类别的概率，具体为图 2 中的输出层：

$$y=Soft\,\mathrm{max}(\boldsymbol{W}_i\boldsymbol{U}_i), \tag{20}$$

最终得到的 y 表示跨境民族文化类别的概率分布，$\boldsymbol{W}_i\in R^{4de\times L}$ 表示可训练的权重向量。

4 实验

4.1 实验数据集

模型训练数据集主要包括 4 个跨境民族(傣族、泰族(泰国)、彝族、傈僳族(越南))，共选取 39 450 条数据作为训练集，2 144 条数据作为测试集。其中每个类别的数据的数量为 1 000～1 500 条。其中标题的长度大多集中在 10 到 20 个字符，正文的长度在 100 到 250 个字符之间，正文中的句子为 5 到 10 句，跨境民族文化数据选取的类别如表 5 所示，其中 NA 表示文本不属于任何一个类型。

4.2 参数设置及评价指标

本文将标题的长度设置为 20，正文句子长度设置为 35，正文句子个数设置为 6，采用 Gensim 工具包中的 Word2Vec 模型训练词向量，词向量的维度为 100 维，采用 Adam 算法作为加快模型训练速度的优化器，将实验参数学习率设为 0.02，批次处理大小设为 16；训练轮次设置为 20。本文的评价指标主要采用准确率(Acc.)、精确率(P)、召回率(R)和 $F1$ 值。

4.3 实验结果及分析

为了验证模型的性能，文本分别设置了四组实验：基线模型对比实验、消融实验、领域分

表 5 数据集类别设置

Table 5 Settings for category of data set

民族	类别
傣族	傣族丧葬文化、傣族乐器文化、傣族婚姻文化、傣族宗教文化、傣族建筑文化、傣族手工艺术文化、傣族服饰文化、傣族舞蹈文化、傣族节日文化、傣族饮食文化
傈僳族	傈僳族丧葬文化、傈僳族婚姻文化、傈僳族宗教文化、傈僳族服饰文化、傈僳族舞蹈文化、傈僳族节日文化
彝族	彝族节日文化、彝族乐器文化、彝族舞蹈文化、彝族宗教文化、彝族建筑文化、彝族服饰文化
泰族	泰族节日文化、泰族婚姻文化、泰族宗教文化、泰族建筑文化、泰族服饰文化
NA	NA

词对模型性能的影响以及Dropout_rate对模型性能的影响。

实验一　基线模型对比实验

为了验证本文模型的有效性,该实验将6个基线模型和本文模型在仅正文和标题联合正文两种情况下进行训练,实验结果对比如表6所示。

(1)DPCNN[16]:Johnson等提出的一种新型的CNN结构,具有提取远程依赖信息和计算复杂度不高的特性。

(2)FastText[17]:具有快速训练文本和快速预测文本的特点。

(3)TextCNN[9]:Kim等提出的面向文本分类的卷积神经网络,能更好捕捉局部特征。

(4)TextRCNN[18]:Lai等提出的使用一个双向递归网络层和一个池化层来提取文本特征信息。

(5)Bert[19]:Google发布的语言表示模型,具有更加高效、能捕捉更长距离依赖的特性。

(6)HAN[3]:Yang等提出的用于文档集分类的分层注意力网络。

表6　本文方法与基线模型方法的对比

Table 6　Comparison between the method of paper and the baseline model methods

数据集	仅正文				标题联合正文			
评价指标	ACC /%	*P* /%	*R* /%	*F*1 /%	ACC /%	*P* /%	*R* /%	*F*1 /%
DPCNN	76.3	77.1	76.3	76.1	73.4	74.2	73.4	72.7
FastText	74.9	75.5	74.9	73.9	75.6	76.1	75.6	74.7
TextCNN	83.6	73.5	70.0	70.5	87.8	76.5	73.4	74.0
TextRCNN	77.7	77.5	77.7	76.9	77.1	77.4	77.1	76.3
Bert	77.6	78.4	77.6	77.0	78.5	79.2	78.5	77.8
HAN	78.0	70.8	71.8	71.7	84.1	79.1	75.7	75.0
平均值	78.0	75.4	74.7	74.4	79.4	77.0	75.6	75.1
本文模型	**81.2**	**71.6**	**76.6**	**72.6**	**92.4**	**87.3**	**88.0**	**86.9**

由表6可知,6个基线模型在数据集为仅正文和标题联合正文两种情况下均取得了合理的实验结果,其中本文模型的性能相比基线模型更加突出。表6中的大部分基线模型在仅正文的情况下的性能都没有超过标题联合正文的性能,但DPCNN、TextRCNN模型在仅正文情况下*F*1值分别达到了76.1%、76.9%,在标题联合正文的情况下*F*1值分别为72.7%、76.3%,下降了3.4%、0.6%,说明了将正文和标题进行简单的拼接操作,并不能很好的利用两者之间的关系,反而给正文增加了冗余信息,影响模型的性能。

在仅正文的情况下,本文模型的ACC值为81.2%,*F*1值为72.6%,模型效果不明显,在标题联合正文的情况下*F*1值提升了14.3%,该结果证明了本文方法能够充分利用标题和正文之间的联系,以及将标题和正文的联合输入对于模型性能的提升具有很好的效果。相比于Bert这种大型的预训练模型,本文的方法在标题联合正文的情况下ACC值提高了13.9%,*F*1值提高了9.1%,验证了本文方法可以更好地适应跨境民族文化分类工作。

值得注意的是,基线模型和本文模型在标题联合正文的实验结果均高于仅正文输入的实验结果。此外,本文方法在标题联合正文的情况下Acc、*P*、*R*、*F*1值都相对较高,平均增加了11.87%,该实验结果证明了标题具有辅助正文分类的特性,以及将标题和正文联合这一观点的合理性。

实验二　消融实验

为了了解本文提出的各个模块对分类所做的贡献,第二组实验采用本文提出的模型的不同变体来进行验证,以下是不同的模型变体设置。

(1)-masked-self-attention:表示在输入层中去除掉正文和标题的特征提取,将提取词级特征的掩码自注意力机制换成简单的注意力模型。

(2)-融合标题和正文的特征提取层:表示未使用融合标题和正文的句子级特征提取,将输入层中获得的标题和正文的特征编码向量拼接后输送到上下文融合层中。

(3)-上下文特征融合层:表示未使用BiGRU进行上下文特征的提取,将提取到的特征信息输送到输出层进行分类。

(4)-TransE:表示未使用跨境民族文化知识图谱进行实体语义的扩展。

由表7可知,在删掉模型中的某一层时,本文方法的ACC、*P*、*R*和*F*1值均有所下降。在"(-)融合标题和正文的特征提取层"和"(-)上下文特征融合层"两种情况下,相比于本文模

表7 消融实验

Table 7 Ablation experiment

数据集	标题联合正文			
评价指标	ACC(%)	*P*	*R*	*F*1
−masked-self-attention(词级)	75.3	72.5	67.4	68.2
−融合标题和正文的特征提取层	90.2	84.7	82.4	81.7
−上下文特征融合层	88.5	76.8	78.5	77.0
−TransE	90.9	87.2	86.3	85.3
本文模型	**92.4**	**87.3**	**88.0**	**86.9**

型实验结果的ACC、*P*、*R*和*F*1值略微下降,由此表明了“融合标题和正文的特征提取层”在文中捕捉句间关系的能力以及“上下文特征融合层”整合上下文特征信息的优势,在本文模型中起到了关键的作用。特别地,“(−)masked-self-attention(词级)”中ACC、*P*、*R*、*F*1值下降最为明显,分别下降了17.1%、14.7%、20.6%和18.7%,这种现象的出现表明了针对跨境民族文化的语义环境复杂问题,仅采用简单的注意力机制无法关注到上下文特征。

此外,在“(−)TransE”情况下,实验删去了对正文内容进行跨境民族实体语义的增强,实验结果显示ACC、*P*、*R*、*F*1值均有所下降,但相比于基线模型,本文提出的模型在不使用实体语义增强的情况下仍然具有较好的性能,此现象支持了本文提出的方法在特征提取方面的优势以及跨境民族文化知识图谱的融入能够增强实体语义这一结论。

实验三 领域分词对模型性能的影响

为了验证领域分词对模型性能的影响,本实验分别采用通用分词(即:仅采用jieba分词进行分词)和领域分词(即:采用跨境民族文化领域词库+jieba分词进行分词)进行对比实验,结果如表8所示。

表8 领域分词对实验结果的影响

Table 8 Influence of domain word segmentation on experimental results

分词方式	ACC/%	*P*/%	*R*/%	*F*1/%
通用分词	74.2	71.0	64.7	64.8
领域分词	**92.4**	**87.3**	**88.0**	**86.9**

通过表8可知,在跨境民族文化文本分类任务中采用领域分词的模型性能明显优于直接使用jieba分词的分类效果。这是因为跨境民族文本数据中存在大量的专业名词,单纯地采用jieba分词往往达不到理想的效果,如:**“香竹糯米饭”**采用jieba分词可能得到**“香竹 糯米饭”**两个词,然而**“香竹糯米饭”**是专业名词,是不可分的。因此,采用领域分词可以更好地辅助模型进行跨境民族文化特征信息的提取,从而提高分类性能。

实验四 Dropout_rate对模型性能的影响

本实验通过设置不同的Dropout_rate参数进行实验以找到最适合本文模型的参数Dropout_rate值,实验结果如图3所示。

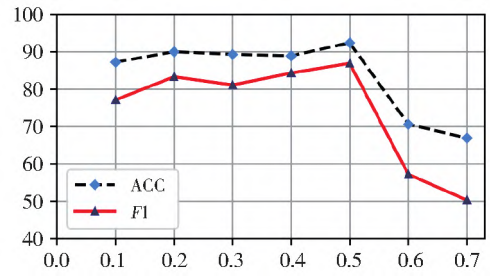

图3 Dropout_rate对模型性能的影响

Fig. 3 The impact of Dropout_rate on model performance

由图3中可以看出,当Dropout_rate=0.7时模型性能最低,*F*1值为50.3%,ACC值为66.9%,原因是Dropout_rate设置得过高时,模型学习到的特征信息较少,使得模型性能明显降低。当Dropout_rate=0.1时,*F*1值为77.1%,ACC值为87.2%,原因是Dropout_rate设置得过低时,模型学习到的信息量较大,导致模型出现过拟合现象。

5 结论

本文针对跨境民族文化文本分类中面临类别交叉的问题,提出了融合领域知识图谱的跨境民族文本分类方法,利用跨境民族文化知识图谱对文本中的跨境民族实体进行语义扩展,通过实体在知识图谱中的类别特征来增强文本的类别语义特征。在此基础上有效利用标题辅助正文锁定关键词、补充和概括正文的优势将其与正文进行联合,并把提取到的不同层次的特征信息结合到一起提升分类的效果,实验结果表明本文提出的方法在跨境民族文化领域相对基线模型具有更好的性能。下一步可考虑将该方法应用于更细粒度的跨境民族文化分类任务。

参考文献:

[1] 崔海亮.“一带一路”背景下中国跨境民族的中华民族认同[J]. 云南民族大学学报(哲学社会科学版), 2016, **33**(1): 35-41. DOI:10. 13727/j. cnki. 53-1191/c. 2016. 01. 007.
CUI H L. Construction of the Chinese National Identity for the Chinese Cross-border Ethnic Groups in the Perspective of “one Belt, one Road”[J]. *J Yunnan Minzu Univ Soc Sci*, 2016, **33**(1): 35−41. DOI:10. 13727/j. cnki. 53 - 1191/c. 2016. 01. 007.

[2] 车蕾, 杨小平, 王良, 等. 面向文本结构的混合分层注意力网络的话题归类[J]. 中文信息学报, 2019, **33**(5): 93−102 DOI:10. 3969/j. issn. 1003-0077. 2019. 05. 011.
CHE L, YANG X P, WANG L, *et al*. Text Structure Oriented Hybrid Hierarchical Attention Networks for Topic Classification[J]. *J Chin Inf Process*, 2019, **33**(5): 93−102 DOI:10. 3969/j. issn. 1003-0077. 2019. 05. 011.

[3] YANG Z C, YANG D Y, DYER C, *et al*. Hierarchical Attention Networks for Document Classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 1480−1489. DOI: 10. 18653/v1/n16-1174.

[4] SHEN T, ZHOU T Y, LONG G D, *et al*. DiSAN: Directional Self-attention Network for RNN/CNN-free Language Understanding[EB/OL]. 2017, arXiv Preprint: 1709. 04696 [cs. CL].

[5] BORDES A, USUNIER N, GARCIA-DURAN A, *et al*. Translating Embeddings for Modeling Multi-relational Data [C]//Neural Information Processing Systems (NIPS), 2013: 1−9.

[6] CHEN J N, HUANG H K, TIAN S F, *et al*. Feature Selection for Text Classification with Naïve Bayes[J]. *Expert Syst Appl*, 2009, **36**(3): 5432−5435. DOI:10. 1016/j. eswa. 2008. 06. 054.

[7] HADDOUD M, MOKHTARI A, LECROQ T, *et al*. Combining Supervised Term-weighting Metrics for SVM Text Classification with Extended Term Representation[J]. *Knowl Inf Syst*, 2016, **49**(3): 909−931. DOI: 10. 1007/s10115-016-0924-1.

[8] ISWARYA P, RADHA V. Ensemble Learning Approach in Improved K Nearest Neighbor Algorithm for Text Categorization[C]//2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015: 1−5. DOI:10. 1109/ICIIECS. 2015. 7193250.

[9] KIM Y. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. DOI:10. 3115/v1/d14-1181.

[10] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A Convolutional Neural Network for Modelling Sentences[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: *Long Papers*), 2014: 655−665. DOI: 10. 3115/v1/P14-1062.

[11] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015: 649−657.

[12] BAHDANAU D, CHO K H, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[C]// The 3rd International Conference on Learning Representations, ICLR 2015.

[13] WANG Y L, WANG S H, TANG J L, *et al*. Hierarchical Attention Network for Action Recognition in Videos[EB/OL]. 2016, arXiv Preprint: 1607. 06416[cs. CV].

[14] LIN Z H, FENG M W, SANTOS C N D, *et al*. A Structured Self-attentive Sentence Embedding[EB/OL]. 2017, arXiv Preprint: 1703. 03130[cs. CL].

[15] 孙新, 唐正, 赵永妍, 等. 基于层次混合注意力机制的文本分类模型[J]. 中文信息学报, 2021, **35**(2): 69−77. DOI: 10. 3969/j. issn. 1003-0077. 2021. 02. 007.
SUN X, TANG Z, ZHAO Y Y, *et al*. Hierarchical Networks with Mixed Attention for Text Classification[J]. *J Chin Inf Process*, 2021, **35**(2): 69−77. DOI:10. 3969/j. issn. 1003-0077. 2021. 02. 007.

[16] JOHNSON R, ZHANG T. Deep Pyramid Convolutional Neural Networks for Text Categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 562−570. DOI:10. 18653/v1/p17-1052.

[17] JOULIN A, GRAVE E, BOJANOWSKI P, *et al*. Bag of Tricks for Efficient Text Classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017: 427. DOI:10. 18653/v1/e17-2068.

[18] LAI S W, XU L H, LIU K, *et al*. Recurrent Convolutional Neural Networks for Text Classification[C]//AAAI'15: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015: 2267−2273.

[19] DEVLIN J, CHANG M W, LEE K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, **1**: 4171−4186. DOI: 10. 18653/v1/N19-1423.