DOI:10.20009/j.cnki.21-1106/TP.2020-1048
2022 年 5 月 第 5 期
Vol.43 No.5 2022

融合领域知识图谱的跨境民族文化分类

毛存礼[1,2],王 斌[1,2],雷雄丽[2,3],满志博[1,2],王红斌[1,2],张亚飞[1,2]

[1](昆明理工大学 信息工程与自动化学院,昆明 650000)
[2](昆明理工大学 云南省人工智能重点实验室,昆明 650000)
[3](昆明冶金高等专科学校,昆明 650000)
E-mail:735361482@qq.com

摘 要: 跨境民族是指居住地"跨越"了国境线,但又保留了原来共同的某些民族特色,彼此有着同一民族的认同感的民族,对于跨境民族文化中涉及到的文本分类问题可以看作领域文本细分类任务,但是,目前面临类别标签歧义的问题.为此提出一种融合领域知识图谱的跨境民族文化分类方法.首先把知识图谱中的知识三元组通过 TransE 模型表示为实体语义向量,并且把实体语义向量与 BERT 预训练模型得到文本中的词语向量相融合得到增强后的文本语义表达,输入到 BiGRU 神经网络中进行深层语义特征提取;然后通过构建注意力权重矩阵,对特征进行权重分配,以此来提升特征的质量,最终完成跨境民族文化分类模型的训练.实验结果表明,提出的方法在跨境民族文化文本数据集上的 F1 值为 89.6%,精确率和召回率分别为 88.2%和 90.1%.

关 键 词: 知识表示;BiGRU;向量融合;TransE;BERT 预训练模型;跨境民族文化分类
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2022)05-0943-07

Cross-border Ethnic Cultural Classification Integrating Domain Knowledge Map

MAO Cun-li[1,2],WANG Bin[1,2],LEI Xiong-li[2,3],MAN Zhi-bo[1,2],WANG Hong-bin[1,2],ZHANG Ya-fei[1,2]

[1](Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650000,China)
[2](Yunnan Key Laboratory of Artificial Intelligence,Kunming University of Science and Technology,Kunming 650000,China)
[3](Kunming Metallurgical College,Kunming 650000,China)

Abstract: Cross-border ethnic cultural classification can be regarded as a sub-category task of domain text,but it faces the problem of category label ambiguity. This article proposes a cross-border ethnic culture classification method based on domain knowledge map. First,the knowledge triples in the knowledge graph are expressed as entity semantic vectors through the TransE model,and the entity semantic vectors are combined with the word vectors in the text obtained by the BERT pre-training model to obtain an enhanced text semantic expression,which is input into the BiGRU neural network Perform deep semantic feature extraction in the middle; then build an attention weight matrix to assign weights to features to improve the quality of features,and finally complete the training of cross-border ethnic cultural classification models. The experimental results show that the F1 value of the proposed method on the cross-border ethnic cultural text data set is 89.6%,and the accuracy and recall rates are 88.2% and 90.1%,respectively.

Key words: knowledge representation;BiGRU;vector fusion;TransE;BERT pre-training model;cross-border ethnic cultural text classification

1 引 言

采用文本分类技术从互联网中获取与跨境民族文化相关的数据,并自动标注所属文化类别,这对开展跨境民族文化融合研究[1]具有重要的价值.在跨境民族文化的文本分类问题中,如何解决标签歧义是当前需要解决的重要问题,例如,文本 1"傣族有很多的节日文化,比如浴佛、丢包、赛龙船等活动"和文本 2"傣族清晨男女老少沐浴更衣到佛寺进行浴佛活动,有些寺院的浴佛方法还是与它的规定有所不同,大致说来这些寺院浴佛更侧重于法会的仪规,具体分为 4 个步骤来进行……"中都含有相同的头实体和尾实体["傣族","浴佛"],但是,尾实体表示的含义又不相同,文本 1 表示的是傣族节日的活动,而文本 2 中所表示的就是傣族宗教的活动.文本 1 中的"浴佛"在知识图谱中的标签为{"傣族","节日","活动"},文本 2 中的"浴佛"在知识图谱中的标签为{"傣族","宗教","活动"},由此可以看出,尾实体产生了歧义的现象,会导致分类错误.

文本分类主流方法主要分为传统机器学习分类算法模型和深度学习神经网络分类算法模型[2].1)基于传统机器学习分类算法模型的核心是利用概率统计的思想对文本中的特征词语进行加权,选择权值较高的词语作为文本特征,以此来进行分类模型的学习[3-6].这类基于特征工程的方法严重依赖

收稿日期:2020-12-15 收修改稿日期:2021-04-05 基金项目:国家自然科学基金项目(61662041,61866019)资助;云南省自然科学基金重点项目(2019FA023)资助;云南省中青年学术和技术带头人后备人才合同项目(2019HB006)资助. 作者简介:**毛存礼**,男,1977 年生,博士,教授,CCF 会员,研究方向为自然语言处理、信息检索、机器翻译;**王 斌**,男,1994 年生,硕士,研究方向为自然语言处理;**雷雄丽**(通讯作者),女,1978 年生,硕士,讲师,研究方向为民族心理学、自然语言处理;**满志博**,男,1996 年生,硕士,CCF 会员,研究方向为自然语言处理、机器翻译;**王红斌**,男,1984 年生,博士,副教授,CCF 会员,研究方向为自然语言处理;**张亚飞**,女,1981 年生,博士,副教授,CCF 会员,研究方向为自然语言处理.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

于人工选取特征的质量,而且很难获取到文本深层的语义特征;2)深度学习是当前文本分类的主流方法,其核心是将文本中的词语以向量的形式进行表示,通过不断的调整网络参数,使输出的数据能够更好的代表输入数据,使用最后的输出作为文本特征进行学习,以此来得到文本的分类模型,如,Keeling 等人[7]提出将卷积神经网络用于法律文献检索任务. 肖琳等人[8]提出一种基于标签语义注意力的多标签文本分类方法. Peng 等人[9]提出了一种新的层次分类意识和注意图胶囊递归 CNNs 框架,用于大规模多标签文本分类. Banerjee 等人[10]提出一种基于层次迁移学习的多标签文本分类算法. 顾天飞等人[11]基于配对排序损失的文本多标签学习算法. Yao 等人[12]提出一种基于图卷积神经网络的文本分类方法. 以上的方法虽然给跨境民族文化分类任务提供了较好的思路,基于传统的机器学习分类的算法模型依赖于数据标注的准确性,针对于跨境民族文化文本分类的问题,数据稀缺并且存在一定的歧义,仅仅利用机器学习的思想无法准确的对跨境民族文本文化进行准确的分类. 基于深度学习的方式是一种数据驱动的方法,需要大规模的分类数据,跨境民族文本文化的数据大多来自于网络,这部分数据较难获取,且如何定义跨境民族文化文本分类的标签也是需要考虑的因素之一,结合知识图谱处理跨境民族文化文本分类问题是一种较好的思路,目前,跨境民族文本文化分类问题中面临的挑战主要有:如何将知识图谱信息有效地和跨境民族分类问题结合以及如何解决民族文化标签歧义的问题.

针对以上在跨境民族文化领域分类存在的问题,提出一种融合领域知识图谱的跨境民族文化分类方法,把跨境民族文化知识图谱中的知识三元组以及实体标签利用 TransE 知识表示模型[13,14]进行向量化表示,采用 BERT 预训练模型进行词向量表示,以增强文本的语义表达. 本文的贡献具体如下:

1)构建了跨境民族文化的知识图谱,并将知识图谱引入到文本分类中,融合了实体的语义信息,扩充了语义信息的表达,缓解了由于标签歧义导致的文本分类不准确的问题.

2)基于预训练 BERT 的思想,增强语义信息,将 BERT 的向量表征与知识图谱向量表征进行融合,得到具有实体语义信息表征的向量,进一步将跨境民族文化中实体信息进行增强.

2 融合知识表示的跨境民族文本分类模型

2.1 跨境民族文化分类模型架构

本文提出的模型架构如图 2 所示,包含了以下 5 个部分:

1)数据输入层:把跨境民族文化知识图谱中实体、关系以及实体标签输入到 TransE 模型中;2)BERT 预训练模型层:基于 Transformer 的最后一层输出的向量作为文本的词语向量;3)TransE 实体向量表示层:对输入的实体、关系以及实体标签进行分布式向量表示,然后进行对位融合得到实体语义向量;4)BiGRU 神经网络层:该层的输入为 TransE 模型输出的实体向量和 BERT 预训练模型层输出的词语向量所融合的增强向量,通过双向 GRU 的门结构对每个词的进行筛选,保留下重要的词语特征,以此来提高文本特征的质量;5)输出层:该层是通过注意力机制对 BiGRU 的输出进行注意力加权,并且利用最大池化的思想获取最显著的信息,再经过一个全连接层,最终通过 Softmax 进行归一化,得到待分类的跨境民族文化文本对应每个类别的得分.

2.2 跨境民族知识图谱构建

知识图谱本质上是一种揭示实体之间关系的语义网络. 知识图谱是由(实体,关系,实体)或(实体,属性,属性值)的三元组形式组成的,通过这些三元组之间的相互连接,可以构成网状的知识结构. 本文以人工构建的方式构建了跨境民族文化知识图谱. 具体的类别如表 1 所示.

表 1 跨境民族文化知识图谱类别

Table 1 Classification of cross-border ethnic fultures knowledge map

序号	大类	序号	小类
1	宗教文化类	1	原始宗教文化
		2	佛教文化
2	建筑文化类	3	民居建筑文化
		4	寺庙建筑文化
3	服饰文化类	5	男性服饰文化
		6	女性服饰文化
4	饮食文化类	7	食品文化
		8	饮品文化
5	艺术文化类	9	乐器艺术
		10	舞蹈艺术
6	习俗文化类	11	节日习俗
		12	婚姻习俗
		13	丧葬习俗文化

在确定跨境民族文化的分类体系后,需要根据各个类别来定义与跨境民族文化相关的属性包括实体的名称、别称、描述内容、实体标签以及实体存在的一些特征. 通过定义实体的这些信息,就可以使实体完整的对跨境民族文化进行详细的描述. 如图 1 所示,对于“泼水节”这个实体来说,它的实体标签类别信息即为“傣族”、“傣族习俗文化”、“傣族节日文化”等. 建立实体与实体之间的关系对跨境民族文化领域知识图谱中的知识进行关联整合,使得跨境民族文化知识图谱更加具有表示性以及提高跨境民族文化知识图谱的查询性能. 跨境民族文化领域的实体关系错综复杂,主要可以归纳为:包含、跨境、位置、同属、属性. 最后通过百科词条信息和结构化知识的组合就可以得到知识三元组信息. 具体如图 1 所示.

2.3 基于 TransE 的跨境民族文化知识表示

本文采用 TransE 模型进行实体语义向量表示,将实体、关系以及实体标签信息训练成分布式向量,然后对这 3 种向量进行对位累加得到实体语义向量. 相比于传统的 TransE 模型来说,由于在训练的过程中添加了实体标签信息,所以本文的 TransE 基本计算如公式(1)所示:

$$(\boldsymbol{h}+\boldsymbol{L}_h)+\boldsymbol{r}\approx(\boldsymbol{t}+\boldsymbol{L}_t) \tag{1}$$

在三元组训练的过程中,由于没有明显的监督信号,也就是不会明确告诉模型学到的知识表示是否正确,所以需要根据正确的三元组 $\boldsymbol{S}$ 构造一些错误的三元组 $\boldsymbol{S}'$,其中 $\boldsymbol{S}'$的构造规则为将正确的三元组中的实体、关系或者实体标签随机替换为其它元素. 在模型训练的过程中,通过设置一个损失函数 L 来对这些三元组进行打分,相比之下,正确的三元组打分要高于错误的三元组,损失函数设计如公式(2)所示:

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$L=\sum_{(h,r,t,L)\in S(h',r,t',L')\in S'}\sum\left[\gamma+\left(\left(\boldsymbol{h}+\boldsymbol{L}_h\right)+\boldsymbol{r}-\left(\boldsymbol{t}+\boldsymbol{L}_t\right)\right)^2-\left(\left(\boldsymbol{h}'+\boldsymbol{L}_{h'}\right)+\boldsymbol{r}-\left(\boldsymbol{t}'+\boldsymbol{L}_{t'}\right)\right)^2\right]_+ \tag{2}$$

其中,$\boldsymbol{h}'$和$\boldsymbol{t}'$为随机构造的负例头实体和尾实体,$\boldsymbol{L}_{h'}$和$\boldsymbol{L}_{t'}$为随机构造的负例头实体和尾实体标签,γ为大于0的超参,+的作用是筛选,具体规则为大于0取原值,小于0则为0.

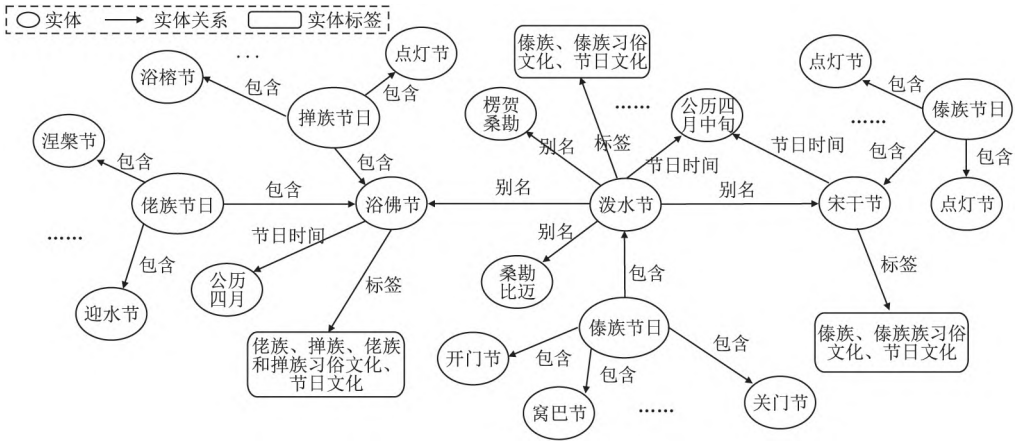

图1　跨境民族文化知识图谱构建示例图

Fig. 1　An example of the construction of a cross-border ethnic cultural knowledge map

训练TransE模型时,首先需要把三元组的实体、关系和该三元组的标签分别按序进行id标记,具体形式为(实体,id)、(关系,id)和(实体标签,id),训练数据格式为(头实体,尾实体,关系,实体标签),模型的输入为随机初始化的实体量、关系向量以及实体标签向量,向量维度一致.通过不断地对实体和关系的向量进行调整,使其满足公式(2)的计算,就可以得到最终的实体向量$\boldsymbol{E}_{id}$和关系向量$\boldsymbol{R}_{id}$和实体标签向量$\boldsymbol{L}_{id}$,把这3种向量进行对位累加得到相应的实体语义向量.

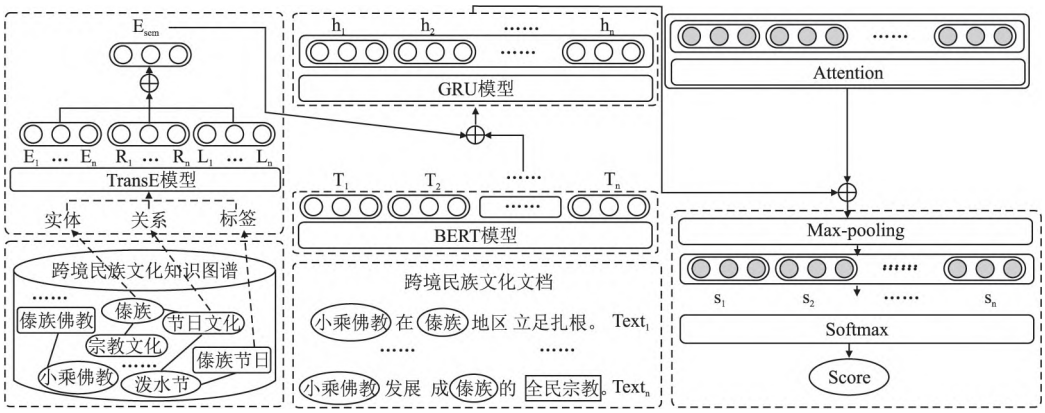

图2　模型构架图

Fig. 2　Model architecture diagram

如图2所示,对于知识三元组[傣族,节日,泼水节],其标签信息为{"傣族","傣族习俗文化","傣族节日"},首先把该三元组的实体、关系及其标签信息进行id标记,得到实体标记为(傣族,0)、(泼水节,1),关系标记为(节日,0),实体"傣族"的标签标记为("少数民族"0_0)、("跨境民族",0_1),实体"泼水节"的标签标记为("傣族习俗文化",1_0)和("傣族节日文化",1_1)然后为每个标记的id初始化相同维度的向量作为TransE模型的输入,通过TransE模型后可以得到实体向量[$\boldsymbol{E}_{傣族}$,$\boldsymbol{E}_{泼水节}$]、关系向量[$\boldsymbol{R}_{节日}$]和相应的实体标签向量[$\boldsymbol{L}_{少数民族}$,$\boldsymbol{L}_{跨境民族}$,$\boldsymbol{L}_{傣族习俗文化}$,$\boldsymbol{L}_{傣族节日文化}$],再把相应的实体标签向量与关系向量对位累加到实体向量中就可以得到实体语义向量,实体"傣族"的语义向量$\boldsymbol{E}^{sem}_{傣族}=\boldsymbol{E}_{傣族}\oplus\boldsymbol{R}_{节日}\oplus\boldsymbol{L}_{少数民族}\oplus\boldsymbol{L}_{跨境民族}$,同理可得实体"泼水节"的语义向量.

$$\boldsymbol{E}^{sem}_{泼水节}=\boldsymbol{E}_{泼水节}\oplus\boldsymbol{R}_{节日}\oplus\boldsymbol{L}_{傣族习俗文化}\oplus\boldsymbol{L}_{傣族节日文化}$$

2.4　基于BERT预训练的文本词向量表征

BERT通过双向Transformer对文本进行表征,在模型处理某一个词语时,如:句子"香茅草烤鱼是傣族的传统美食",分词之后可以得到[香茅草烤鱼是傣族的传统美食],BERT模型会随机遮罩一些词汇得到"香茅草烤鱼是[Mask]的传统美食",然后根据上下文信息对[Mask]进行预测,这样就可以很好的把上下文的语义信息融入到[Mask]这个词语的表示中中句子"香茅草烤鱼是傣族的传统美食"为文本句子,"傣族,傣族饮食文化,傣族食品"为文本中的实体对[香茅草烤鱼,傣族]的标签信息,"傣族,傣族菜,香茅草烤鱼"为知识图谱中的三元组信息. Transformer Encoder的输入Input Embedding为文本经过Token Embedding,Segment Embedding和

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Position Embedding 后按位相加的词语向量,例如文本“泼水节是傣族的传统节日”经过以上 3 个 Embedding 的元素按位相加后表示为 $\boldsymbol{A}=\{\boldsymbol{a}_{[CLS]},\boldsymbol{a}_{泼水节},\boldsymbol{a}_{是},\boldsymbol{a}_{傣族},\boldsymbol{a}_{的},\boldsymbol{a}_{传统},\boldsymbol{a}_{节日},\boldsymbol{a}_{[SEP]}\}$,其中 $\boldsymbol{a}_{[CLS]}$ 和 $\boldsymbol{a}_{[SEP]}$ 为文本的特殊标记向量,每个词语都被表示为 k 维的向量. 对于输入的向量利用 Multi-Head Attention(多头注意力机制)计算文本中每个词语与其它词语之间的相互关系,计算公式如公式(3)-公式(5)所示.

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}(\frac{\boldsymbol{QK}^{\boldsymbol{T}}}{\sqrt{\boldsymbol{d}_k}})\boldsymbol{V} \tag{3}$$

$$\boldsymbol{MHA}=\mathit{Concat}(\text{head}_1,\cdots,\text{head}_k)\boldsymbol{W}^{O} \tag{4}$$

$$\text{head}_i=\text{Attention}(\boldsymbol{QW}_i^{\boldsymbol{Q}},\boldsymbol{KW}_i^{\boldsymbol{K}},\boldsymbol{VW}_i^{\boldsymbol{V}}) \tag{5}$$

其中,$\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{W}_i^{\boldsymbol{V}}$ 表示第 $\boldsymbol{i}$ 个 Head 的 $\boldsymbol{W}^{\boldsymbol{Q}},\boldsymbol{W}^{\boldsymbol{K}},\boldsymbol{W}^{\boldsymbol{V}}$ 矩阵,$\boldsymbol{W}^{\boldsymbol{O}}$ 为附加权重矩阵,$\boldsymbol{d}_k$ 为输入词向量的维度.

2.5 融合实体语义向量的词向量表征

本文以融合实体语义向量的方式对文本的词向量进行增强,由于 BERT 预训练模型得到的词向量是包含文本的上下文序列信息的词向量;而 TransE 模型得到的实体向量是包含了头实体向量和尾实体向量,它们之间存在紧密的联系,因此,通过对词向量以及实体向量进行融合可以增加实体在文本中的表示性. 通过 2.3 节和 2.4 节可以分别得到三元组的实体向量表示 $\boldsymbol{E}_{\text{id}}^{\text{sem}}$ 和文本中每个词语的词向量表示 A_{t},并且这两个向量的维度是一致的,通过实体在文本中的位置可以进行这两种向量的对位相加得到文本的词向量 $\boldsymbol{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$,融合过程如公式(6)所示:

$$\boldsymbol{W}=\boldsymbol{A}_t\oplus\boldsymbol{E}_{\text{id}}^{\text{sem}} \tag{6}$$

其中,$\boldsymbol{A}_t$ 为 BERT 输出的每个词的词向量,$\boldsymbol{E}_{\text{id}}^{\text{sem}}$ 为维度与 $\boldsymbol{A}_t$ 一致的三元组实体语义向量. 例如:对于文本“香茅草烤鱼是傣族的传统美食”,其中含有三元组[“傣族”,“傣族菜”,“香茅草烤鱼”],实体“傣族”的标签为{“少数民族”,“跨境民族”},实体“香茅草烤鱼”的标签信息为{“傣族饮食文化”,“傣族食品”},通过 2.3 节的 TransE 知识表示模型,最终可以得到实体语义向量$[\boldsymbol{E}_{傣族}^{\text{sem}},\boldsymbol{E}_{傣香茅草烤鱼族}^{\text{sem}}]$;通过 2.4 节的 BERT 预训练模型可以得到文本中每个词语的向量表示 $\boldsymbol{W}=\{\boldsymbol{w}_{香茅草烤鱼},\boldsymbol{w}_{是},\boldsymbol{w}_{傣族},\boldsymbol{w}_{的},\boldsymbol{w}_{传统},\boldsymbol{w}_{美食}\}$,然后通过词语的 id 就可以把实体向量按位相加到实体的词向量上;最终得到的词向量表示为 $\boldsymbol{W}=\{\boldsymbol{w}_{香茅草烤鱼}+\boldsymbol{E}_{傣香茅草烤鱼族}^{\text{sem}},\boldsymbol{w}_{是},\boldsymbol{w}_{傣族}+\boldsymbol{E}_{傣族}^{\text{sem}},\boldsymbol{w}_{的},\boldsymbol{w}_{传统},\boldsymbol{w}_{美食}\}$,通过融合后就可以把实体“香茅草烤鱼”与实体“傣族”之间存在的相互联系加入到文本的语义特征中.

2.6 基于 BiGRU 神经网络的文本特征抽取

GRU 是 Chung 等人[15]提出的 LSTM 的一个变种,既继承了 LSTM 可以学习长期依赖信息的特性,而且又减少了训练参数,提高了模型的训练效率. BiGRU 的输入 $\boldsymbol{x}$ 的表示如公式(7)所示:

$$\boldsymbol{x}_i=\{\boldsymbol{w}_i+\boldsymbol{E}_i,\boldsymbol{p}_1,\boldsymbol{p}_2\} \tag{7}$$

其中,$\boldsymbol{p}_1$ 表示第这个词语与第 1 个实体“香茅草烤鱼”和第 2 个实体“傣族”之间的位置向量,因为该词语就是第 1 个实体本身,相对位置的 id 为 0,所以 $\boldsymbol{p}_1$ 的值为与词向量维度相同的随机初始化向量,同理可知该词语到第 2 个实体的相对位置的 id 为 2,所以 $\boldsymbol{p}_2$ 的值为与词向量维度相同的随机初始化向量.

在 BiGRU 中,新的记忆 $\tilde{\boldsymbol{h}}_1$ 是由过去的隐含状态 $\tilde{\boldsymbol{h}}_{i-1}$ 和新的输入 $\tilde{\boldsymbol{x}}_i$ 决定的:

$$\tilde{\boldsymbol{h}}_1=\tanh(\boldsymbol{W}_{\tilde{h}}\cdot[\boldsymbol{r}_i\times\boldsymbol{h}_{i-1},\boldsymbol{x}_i]) \tag{8}$$

其中,tanh( )是激活函数,$\boldsymbol{h}_{i-1}$ 为上一个时刻的隐含状态,r_i 是重置信号,它用来判定上一个隐含状态 $\boldsymbol{h}_{i-1}$ 对结果 $\tilde{\boldsymbol{h}}_1$ 的重要程度.

$$\boldsymbol{r}_i=\sigma(\boldsymbol{W}_r\cdot[\boldsymbol{x}_i,\boldsymbol{h}_{i-1}]) \tag{9}$$

其中,σ( )是激活函数 Sigmoid 函数,其值域范围在(0,1)之间.

更新门 $\boldsymbol{z}$ 决定的是上一个隐含状态 $\boldsymbol{h}_{i-1}$ 向下一个状态传递的信息. 控制 $\boldsymbol{h}_{i-1}$ 中有多少信息可以流入 $\boldsymbol{h}_i$ 中.

$$\boldsymbol{z}=\boldsymbol{\sigma}(\boldsymbol{W}_z\cdot[\boldsymbol{x}_i,\boldsymbol{h}_{i\text{-}1}]) \tag{10}$$

隐含状态 $\boldsymbol{h}_i$ 由上一个隐含状态 $\boldsymbol{h}_{i-1}$ 产生,新的记忆由更新门判定.

$$\boldsymbol{h}_i=(1-z)\cdot\tilde{\boldsymbol{h}}_1+\text{z}\cdot\boldsymbol{h}_{\text{i}-1} \tag{11}$$

其中,上述公式中的 $\boldsymbol{W}_{\tilde{h}},\boldsymbol{W}_r,\boldsymbol{W}_z$ 是在训练 GRU 时所学到的参数. 由于本文采用双向的 GRU,以此来获取文本正向和反向的跨境民族文化文档上下文信息,$\boldsymbol{h}_i=[\overrightarrow{\boldsymbol{h}_i}\oplus\overleftarrow{\boldsymbol{h}_i}]$.

2.7 基于 Attention 机制的特征加权

根据对跨境民族文化数据的分析,文本中的某些关键特征词具有很重要的语义信息,需要着重的进行考虑. 因此,本文利用注意力机制来为这些特征词语分配更高的权重,突出这些特征的重要性. 通过 2.6 节可以得到文本中的第 $\boldsymbol{i}$ 个文本特征词语的向量表示 $\boldsymbol{h}_i$,通过随机初始化一个向量 $\boldsymbol{u}_w$ 作为模型参数一起训练,得到每个词语的注意力得分 $\boldsymbol{\alpha}_i$,计算如公式(12)所示:

$$\boldsymbol{\alpha}_i=\frac{\exp(\boldsymbol{h}_i\boldsymbol{u}_W)}{\sum_{i=1}^{n}\exp(\boldsymbol{h}_i\boldsymbol{u}_W)} \tag{12}$$

令第 i 个文本特征词语的向量表示 $\boldsymbol{h}_i$ 与其注意力得分 $\boldsymbol{\alpha}_i$ 相乘,从而获得该词语新的特征向量. 最后采用最大池化的思想获取最显著的跨境民族文化特征信息,计算如公式(13)所示:

$$\boldsymbol{c}_i=\text{maxpool}(\sum_i^T\boldsymbol{h}_i\boldsymbol{\alpha}_i) \tag{13}$$

对于输入的文本来说,通过注意力机制加权后可以得到该句子的向量形式表示 $\boldsymbol{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n\}$,其中 $\boldsymbol{C}\in\boldsymbol{R}^{n\times d}$ 为句子向量,$\boldsymbol{d}$ 为句子向量的维度,$\boldsymbol{n}$ 为文本数据的词语数量. 再经过一个全连接层可以得到输出为 $\boldsymbol{Y}$ 的一维向量,表示为 $\boldsymbol{Y}=[\boldsymbol{y}^1,\boldsymbol{y}^2,\cdots,\boldsymbol{y}^k]$,其中 k 为类别数,$\text{y}^{\text{i}}$ 为输入的句子向量 C 属于第 i 类的预测值,y^{i} 的计算方式如公式(14)所示:

$$y^i=W_i\cdot C+b \tag{14}$$

其中,W_i 为该句子对应类别 i 的权重矩阵,b 为偏置值,表示为 $\text{b}=[b_1,b_2,\cdots,b_k]$. 通过公式(14)得到 y^{i} 后,再通过 Softmax 函数进行归一化处理,得到 C 属于各个类别的概率值,公式如公式(15)所示:

$$p(y=j\mid C)=\mathit{soft}\max(y^j) \tag{15}$$

其中,公式(15)表示句子 C 属于类别 j 的概率值.

2.8 模型训练及优化策略

本文使用交叉熵损失函数作为目标函数,通过刻画预测标签与实际标签之间的距离来判定这两者的接近程度,也就是交叉熵越小,距离越近,预测标签与实际标签越相似. 目标函数定义如公式(16)所示:

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$J(\theta) = -\sum_{i=1}^{T} \log p(y=j|C) \tag{16}$$

其中,θ 表示模型中的所有参数,初始值随机;T 代表句子集合数,本文使用 Adam 优化器对参数进行更新.

3　实　验

3.1　实验数据集

本文所使用的数据集包含两部分:

1)跨境民族文化知识图谱:其中包括了 863 个三元组,13 个小类. 其中知识三元组的具体格式是["实体","关系","实体"]或者["实体","属性","属性值"],例如:知识三元组["傣族","节日","泼水节"]和["泼水节","时间","公历 4 月 13 ~ 15 日"].

2)文本数据:利用已经构建好的跨境民族文化知识图谱中的知识三元组与跨境民族文化文本进行实体对齐所获取的实验数据. 如果知识图谱中三元组的头实体和尾实体同时出现在跨境民族文化文本中,我们就把这个文本归为实验所需的标注数据,对于这些标注好的数据则利用人工进行校验,然后对每条数据打上类别标签. 标注数据的格式为:[标签- > 文本]. 本文实验从跨境民族文化领域文本集中抽取了 40 种类别共计 46251 条语料,4000 条作为测试集,标注的每条数据的平均长度为 67 个字符,总共标注的类别有 40 个. 每个类别的数据的数量为 1110 ~ 1190 条. 而且本次实验中还加入了一些特殊的文本类别 NA(NA:表示句子不属于任何一个文本类别),实验数据示例如表 2 所示.

表 2　标注数据样例

Table 2　Sample label data

傣族食品	到过云南西双版纳的人,几乎无人会忘记香茅草烤鱼的美味,在傣味菜肴中知名度非常高,是傣族款待贵宾不可少的一道菜,在泼水节的时候家家户户都有这道菜出现在餐桌上.
傣族食品	香茅草烤鱼是地道传统的一道傣族风味菜肴.
……	……

3.2　实验参数设置

实验过程中,通过不断的调节实验参数,以确保模型在参数最优的情况下进行训练,具体的参数设置如表 3 所示.

表 3　模型参数设置

Table 3　Model parameter setting

Object	Number
Net_model	ATT_BiGRU
Layers	2
Optimizer	Adam
Max_len	150
Vocab_size	33693
Word vector	100
Num_epochs	20
Num_classes	40
Dropout_rate[16]	0.5
Batch_size	20
Learnning_rate	0.01

3.3　实验评测指标

本文为了证明实验的有效性,通过精确率(Precision)、召回率(Recall)和 F_1 值来对模型进行评估. 精确率、召回率和 F_1 值的计算方法如公式(17)-公式(19)所示.

$$\text{Precision} = \frac{\text{Right\_num}}{\text{Recognize\_num}} \tag{17}$$

$$\text{Recall} = \frac{\text{Right\_num}}{\text{All\_num}} \tag{18}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

其中,Right_num 为预测正确的文本数量,Recognize_num 为识别出的文本数量,All_num 为此次测试的文本数量. 由于本文的任务是做跨境民族文化文本分类任务,需要在保持高精确率的情况下有一个高召回率,所以 F1 越高代表模型的平衡性越好,分类效果越好.

3.4　实验结果与分析

实验 1. 不同方法实验结果对比

为了验证本文方法的有效性,在相同实验语料的情况下,设计了 7 组不同分类方法的对比实验进行本文方法有效性的验证. 其中,各个模型的实验数据完全一致,实验中使用领域分词的方法对文本进行预处理.

1)文献[17]所提出的一种基于 word-level 级别的深层卷积神经网络模型 DPCNN 文本分类模型;

2)文献[18]所提出的基于 Attention_BiLSTM 的神经网络文本分类方法;

3)文献[19]提出的 TextCNN 文本分类经典模型;

4)文献[20]所提出的 Transformer 模型应用于文本分类的方法;

5)文献[21]提出的 BiLSTM-CNN 文本分类模型;

6)文献[22]提出的 FastText 文本分类模型. 实验结果如表 3 所示;

7)Baseline(Attention_BiGRU):Attention 是指注意力机制,这一机制已经被广泛应用于多种领域,包括图像标题生成、文本分类、语音识别和机器翻译[24]. 双向门控循环神经网络(BiGRU)可以看做双向长短时记忆网络(BiLSTM)的一种拓展[23].

如表 4 所示,本文方法在跨境民族文化文本分类任务上的精确率和召回率方面都优于本文的 Baseline 以及其他方法.

表 4　与其它模型的分类效果对比

Table 4　Compared with the classification effect of other models

模型名称	P(%)	R(%)	F1(%)
DPCNN	0.865	0.873	0.857
Attention_BiLSTM	0.859	0.864	0.852
TextCNN	0.853	0.859	0.846
Transformer	0.884	0.867	0.859
BiLSTM-CNN	0.888	0.884	0.876
FastText	0.873	0.866	0.858
Attention_BiGRU	0.857	0.869	0.863
本文方法	**0.882**	**0.901**	**0.896**

对于本文的 Baseline 方法 Attention_BiGRU 来说,本文方法优于它的原因是本文的词向量表示使用的是 BERT 模型,所表示的每个词语都带有上下文语义信息,使特征更具有代

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表性. 而 Baseline 方法的词向量表示使用的是 Word2vec 模型,而且还没有融入实体向量和对特征进行加权. 所以本文方法优于 Baseline 模型 Attention_BiGRU. Transformer 模型和 BiLSTM-CNN 模型的精确率优于本文的模型,造成这种结果的原因是这两个网络模型的网络层数大于本文模型的网络层数. 而对于网络层数更深的 DPCNN 模型来说,其结果不理想的原因是因为网络模型单一,而且词语级的输入不能很好的对文本进行表示.

实验 2. 不同词向量表示方法对实验结果的影响

为了验证本文所使用的 BERT 预训练模型表示的文本词向量对于分类任务的有效性. 本文通过几种不同的向量表征方式来对文本进行表征,其中的详细实验方式是分别利用 Word2vec 模型和 Glove 模型对文本进行词向量表示,并且与 TransE 模型的实体向量进行融合,而其它保持不变进行模型训练. 实验结果如表 4 所示.

表 5 不同词向量方式对实验结果的影响

Table 5 Influence of different word vector modes on experimental results

词向量方法	P(%)	R(%)	F1(%)
Word2vec	0.864	0.882	0.877
GloVe	0.871	0.889	0.884
BERT	**0.882**	**0.901**	**0.896**

从表 5 可以看出,本文通过把 BERT 预训练模型所表示的文本词向量和 TransE 模型所表示的实体向量进行融合,在跨境民族文化文本分类任务上具有较好的性能. 其根本原因在于 BERT 预训练模型对文本中的词语进行向量表示时,利用双向 Transformer 对文本中的每个词语进行表示,充分考虑了文本的上下文语义信息;而 Word2vec 模型只考虑了词语的局部信息,没有考虑词语与局部窗口之外词的联系;GloVe 模型虽然弥补 Word2vec 模型的缺陷,考虑了词语的整体信息,但还存在一个问题,就是所表示的词语在不同语境下的词向量是相同的,没有考虑语境的问题;BERT 模型对于上述问题都进行了综合的考虑,即考虑了词语的局部以及整体信息,又考虑了词语在不同语境下的词向量变化,能够充分的对文本中的每个词语进行表示.

实验 3. 领域词汇对实验结果的影响

由于本文需要通过融入领域实体来解决文本中实体特征存在歧义的问题. 本文通过领域分词的方法来对文本进行分词处理,以此来保证文本中实体特征词的完整性. 所以本文分别采用通用分词工具和领域分词分词实验对比,其中,通用分词使用 jieba 分词工具,领域分词采用构建的领域词典 + jieba 分词,实验结果如表 5 所示.

表 6 领域分词对实验结果的影响

Table 6 Influence of domain segmentation on experimental results

分词方式	P(%)	R(%)	F1(%)
通用分词	0.872	0.890	0.885
领域分词	**0.882**	**0.901**	**0.896**

从表 6 可以看出,采用领域分词的效果明显高于直接使用 jieba 分词的效果. 本文中将跨境民族文化相关文本中由多个词汇构成的跨境民族文化特征词汇作为领域词汇来处理,如,“南传上部座佛教”这个词语在使用 jieba 分词时可以分为“南传”、“上部座”和“佛教”这 3 个独立的词语,而利用领域分词就可以得到一个完整的词语. 诸如此类的词语还有很多,如:浅色大襟短衫、大襟小袖短衫. 这些词汇如果直接使用 jieba 分词后将导致具有完整语义的设备缺陷特征拆开后导致语义信息丢失,而作为领域词汇利用 BERT 进行词向量表征后能够有效获取到跟跨境民族文化相关的词汇的语义特征,更有利于通过 Attention 层进行捕捉.

实验 4. 不同实验参数对实验结果的影响

设置不同的参数迭代次数、批次大小以及随机失活率进行实验的结果如图 3-图 5 所示,根据实验结果可知,设置 epochs 为 20、Dropout 为 0.4 以及 Batch 为 32 时,实验结果达到了最佳.

4 结束语

针对跨境民族文化标签类别存在歧义的问题. 本文提出了融合领域知识图谱的跨境民族文化文本分类方法,该方法利用 TransE 知识表示模型得到文本中知识三元组,利用 BERT 预训练模型得到文本中每个词语的向量表示,再通过 BiGRU 神提取文本的深层语义信息. 实验表明,本文方法对于特定领域的文本分类任务有着良好的效果. 在未来的工作中将进一步的解决向量的表征问题和提升文本特征的质量,使模型的抽取效果进一步的提升.

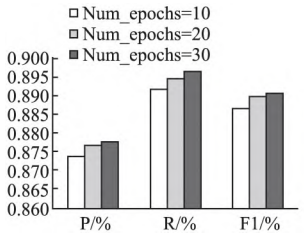

图 3 Num_epochs 对实验结果的影响

Fig. 3 Influence of Num_epochs on experimental results

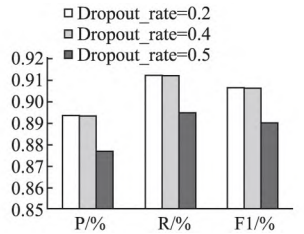

图 4 Dropout_rate 对实验结果的影响

Fig. 4 Influence of Dropout_rate on experimental results

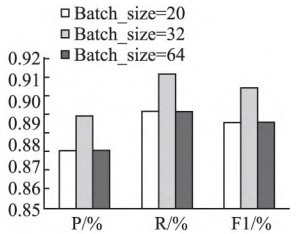

图 5 Batch_size 对实验结果的影响

Fig. 5 Influence of Batch_size on experimental results

References:

[1] Cui Hai-liang. Chinese national identity of cross-border nationalities

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

under the background of " one belt and one road" [J]. Journal of Yunnan University for Nationalities(Philosophy and Social Sciences Edition),2016,33(1):35-41.
[2] Lecunn Y,Bengio Y,Hinton G. Deep learning[J]. Nature,2015,521(7553):436-444.
[3] Chen J,Huang H,Tian S,et al. Feature selection for text classification with nave bayes[J]. Expert Systems with Applications an International Journal,2009,36(3):5432-5435.
[4] Qian Tao,Xiong Hui,Chen Jin-yin. Pearl fine-grain classification method based on MV-PearlNet[J]. Journal of Chinese Computer Systems,2021,42(1):185-190.
[5] Iswarya P,Radha V. Ensemble learning approach in improved K nearest neighbor algorithm for text categorization[C]//Proceedings of the International Conference on Innovations in Information,Embedded and Communication Systems(ICIIECS),Piscataway: Institute of Electrical and Electronics Engineers,2015:1-5.
[6] Haddoud M,Mokhtari A,Lecroq T,et al. Combining supervised term-weighting metrics for SVM text classification with extended term representation [J]. Knowledge and Information Systems,2016,49(3): 909-931.
[7] Keeling R,Chhatwal R,Huber-Fliflet N,et al. Empirical comparisons of CNN with other learning algorithms for text classification in legal document review[C]//IEEE International Conference on Big Data (Big Data),IEEE,2019:2038-2042.
[8] Xiao Lin,Chen Bo-li,Huang Xin,et al. Multiple label text classification based on semantic label attention [J]. Journal of Software,2020,31(4):1079-1089.
[9] Peng H,Li J,Gong Q,et al. Hierarchical taxonomy-aware and attentional graph capsule RCNNs for large-scale multi-label text classification[J]. IEEE Transactions on Knowledge and Data Engineering,2021,33(6):2505-2519.
[10] Banerjee S,Akkaya C,Perez-Sorrosal F,et al. Hierarchical transfer learning for multi-label text classification[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics(ACL),Florence,2019:6295-6300.
[11] Wang Jin,Chen Zhi-liang,Li Hang,et al. Hierarchical multi-label classification using incremental hypernetwork [J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition),2019,31(4):538-549.
[12] Yao L,Mao C,Luo Y. Graph convolutional networks for text classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence(AAAI),2019,33: 7370-7377.
[13] Bordes A,Usunier N,Garcia-Duran A,et al. Translating embeddings for modeling multi-relational data[C]//Advances in Neural Information Processing Systems (NIPS),Lake Tahoe,Nevada,USA. Lake Tahoe,Nevada,December 5th-10th,2013,Stroudsburg: NIPS,2013: 2787-2795.
[14] Cheng Shu-yu,Huang Shu-hua,Yin Jian. Recommendation model based on knowledge graph and recurrent neural network[J]. Journal of Chinese Computer Systems,2020,41(8):1670-1675.
[15] Chung J,Gulcehre C,Cho K H,et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. Arxiv preprint arXiv:1412.3555,2014.
[16] Srivastava N,Hinton G,Krizhevsky A,et al. Dropout:a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research,2014,15(1):1929-1958.
[17] 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(ACL),2017:562-570.
[18] Zhou P,Shi W,Tian J,et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(ACL),2016:207-212.
[19] Kim Y. Convolutional neural networks for sentence classification [J]. arXiv preprint arXiv:1408.5882,2014.
[20] Vaswani A,Shazeer N,Parmar N,et al. Attention is all you need [C]//Advances in Neural Information Processing Systems (NIPS),2017:5998-6008.
[21] Xie P,Wang G,Zhang C,et al. Bidirectional recurrent neural network and convolutional neural network (BiRCNN) for ECG beat classification[C]//40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),2018:2555-2558.
[22] 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 Pages,2017:427-431.
[23] Zhang Q,Ma Y,Gu M,et al. End-to-end Chinese dialects identification in short utterances using CNN-BiGRU[C]//IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC),IEEE,2019:340-344.
[24] Vaswani A,Shazeer N,Parmar N,et al. Attention is all you need [C]//Advances in Neural Information Processing Systems (NIPS),2017:5998-6008.

附中文参考文献:

[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.
[4] 钱　涛,熊　晖,陈晋音. 基于 MV-PearlNet 的珍珠细粒度分类方法[J]. 小型微型计算机系统,2021,42(1): 185-190.
[8] 肖　琳,陈博理,黄　鑫,等. 基于标签语义注意力的多标签文本分类[J]. 软件学报,2020,31(4):1079-1089.
[11] 王　进,陈知良,李　航,等. 一种基于增量式超网络的多标签分类方法[J]. 重庆邮电大学学报(自然科学版),2019,31(4): 538-549.
[14] 程淑玉,黄淑桦,印　鉴. 融合知识图谱与循环神经网络的推荐模型[J]. 小型微型计算机系统,2020,41(8): 1670-1675.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net