融合 La 格虚词语义信息的藏文 La 格分类模型

班玛宝，慈祯嘉措[1,2]，张瑞，才让加[1,2,3,4,5]*

（1. 青海师范大学 计算机学院，青海 西宁 810016；2. 省部共建藏语智能信息处理及应用国家重点实验室，青海 西宁 810008；3. 青海省藏文信息处理与机器翻译重点实验室，青海 西宁 810008；4. 藏文信息处理教育部重点实验室，青海 西宁 810008；5. 青海省藏文信息处理工程技术研究中心，西宁 810008）

摘要：采用深度学习方法实现藏文 La 格(ལ་དོན།)分类是一项具有挑战性和重要研究意义的藏语自然语言处理任务。藏文La格的自动分类更加依赖于上下文语义信息和特征的时序性，该文通过分析La格虚词的语义特征及用法，在设计 La 格虚词语义信息标记算法的基础上，提出一种融合 La 格虚词语义信息的藏文 La 格分类模型。该模型首先以每个音节及对应 La 格虚词或其它音节的语义特征嵌入作为输入，丰富嵌入向量的语义信息，增加输入特征的多样性；然后采用一维卷积融合并学习每个音节及对应 La 格虚词或其它音节语义信息的局部特征向量，提高卷积层的空间特征学习能力；其次使用双向 LSTM 学习时序特征，提高时序特征的学习能力；最后使用注意力机制对双向 LSTM 层每一时刻的输出特征进行加权融合，充分利用每一时刻的输出特征，以提高最终文本表示的特征质量。经在 TLD 数据集上实验显示，该模型的分类效果比基线模型和仅用藏文音节嵌入的模型均有所提升，在测试集上的分类准确率为 93.10%。

关键词：自然语言处理；La 格虚词；语义信息；神经网络；La 格分类

中图分类号：TP391.1　　**文献标志码：**A

La 格是藏文语法典籍《三十颂》中的重点和难点[1]，也是八格（ཇུག་དྲྭ་བརྒྱད་པོ།）中的主要研究内容。传统藏文文法中，从格语法角度出发，对藏文语义进行了一些探讨和研究，为进一步研究奠定了基础。

光 La 格虚词的用法就占据着八格中的三席，分别是业格（ལས་སུ་བྱ་བ།）、为格（དགོས་ཆེད།）和依格（དགོས་ཆེད།），另外同格（དེ་ཉིད།）和时格（ཚེ་སྐབས།）也是 La 格常见的两类用法。因此详细分析 La 格虚词的几种用法，研究藏文 La 格分类技术，在藏语格语法研究、语法功能研究和自然语言理解等藏语自然语言处理任务中具有广泛的应用前景。此外，La 格是藏语文课本中必学的一个重点知识，唯有熟练掌握其概念和用法，才能准确区分藏文 La 句子类型，并进一步深入分析每个句子的实际语义。可见研究基于机器学习方法的藏文 La 格分类技术在 La 格学习中也具备一定的实际应用价值。

近年来，随着深度学习技术的不断成熟[2-3]，卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network, RNN）和一些混合神经网络被逐渐应用于句子分类任务。因循环神经网络存在梯度消失和梯度爆炸问题，所以通常会使用其变体长短时记忆网络（long short-term memory，LSTM）。文献[4]提出了一种基于卷积神经网络的句子分类模型，经在 7 个任务上测试，有 4 个任务都取得了当时最新研究结果。文献[5]提出了一种基于注意力的卷积神经网络句子分类模型，经实验表明，该模型的句子分类性能优于传统 CNN 模型。文献[6]使用卷积神经网络进行多组实验，考查了句子分类模型性能对超参数变化的敏感程度。文献[7]提出了基于稀疏自学习卷积神经网络的句子分类模型，经实验表明，该模型取得了较好的分类效果。文献[8]提出了一种用于句子分类的多粒度注意力门控卷积神经网络，经实验表明，该模型的句子分类准确率比标准 CNN 提高了 3.1%。文献[9]提出了极性转移 LSTM 树结构网络模型，更好地获取情感信息来进行句子分类，经实验表明，该模型在斯坦福情感数据集上的句子分类效果优于 LSTM 和递归神经网络等模型。文献[10]提出了一种用于句子分类的胶囊网络(CapsNets)，经实验表明，该网络的句子分类性能优于 CNN 和 RNN 模型。文献[11]提出了一种基于卷积神经网络和贝叶斯分类器的句子分类模型，经实验表明，该模型优于只使用深度学习的模型或传统的句子分类模型。从上述研究可见，英汉文通过理论和方法的创新，对常用句子分类模型进行优化和改进，已成功将深度学习方法运用到了句子分类任务中，并取得了可喜的研究结果，为进一步研究奠定了基础。

常见的藏文文法中都对藏文句型及分类方法进行了一定的研究，为研究基于深度学习方法的藏文句子自动分类技术奠定了理论基础。有关藏文句子自动分类的研究及文献报道极少。文献[12]提出了一种基于循环卷积神经网络的藏文句型分类方法，经实验，其分类准确率为 85.61%。文献[13]提出了一种基于短语特征的藏文疑问句分类方法，经实验，其分类准确率为 96.98%。文献[14]提出了一种融合双通道音节特征的藏文 La 格自动分类模型，经实验，其分类准确率为 90.26%，取得了较好的分类效果，对藏文 La 格自动分类任务的研究具有开创性意义，但该模型存在两点可改进之处，一是该模型没有学习到以藏文 La 格虚词为核心的上下文语义信息，然而经研究发现，藏文 La 格虚词在不同 La 格句型中的语义功能及使用频度均有所差异[1]；二是该模型采用了二维卷积模式，然而研究表明，一维卷积更适用于序列数

收稿日期：2022- XX-XX　　**录用日期：**2022- XX-XX

基金项目：国家自然科学基金资助项目(61866032,619660316)；青海省重点研发项目（2022-GX-104）

* **通信作者：**zwxxzx@163.com

据，如自然语言处理领域，二维卷积则更适用于计算机视觉和图像处理领域[15-16]。

针对藏文 La 格的分类对以 La 格虚词为核心的局部空间特征的依赖度大、对特征的时序性要求高和文献[14]中模型存在的不足，本文通过分析 La 格虚词的语义特征及用法，在设计藏文 La 格虚词语义信息标记算法的基础上，提出了一种融合 La 格虚词语义信息的藏文 La 格分类模型，该模型以每个音节及对应 La 格虚词或其它音节的语义信息作为输入，增加了特征的多样性，丰富了输入特征的语义信息；为了提高局部空间特征的学习能力，采用一维卷积操作学习局部特征并融合每个音节及对应 La 格虚词或其它音节的语义信息，提高了卷积层的空间特征学习能力；为了提高模型学习时序特征的能力，使用双向 LSTM 对时序特征进行了更合理地学习；为了充分利用双向 LSTM 层各时刻的输出特征，使用注意力机制对每一时刻的输出特征进行加权融合，提高了最终的特征质量；最后，经进行多组对比实验表明，该模型可取得更好的分类效果，验证了本文模型的优越性。

1 藏文 La 格虚词语义信息标记算法

1.1 La 格虚词的用法及语义特征分析

由于 La 格中的虚词“སྤ་ར་ན་ཏུ་ན་དྲ”与虚词“ལ”在业格、为格、依格、同格和时格中的用法基本相同，所以统称它们为 La 格。根据 La 格虚词的语法和语义功能，La 格虚词的用法可以分成表示业格、为格、依格、同格和时格的五类 La 格句型，藏文 La 格句子实例见表 1.

表 1 藏文 La 格句子实例

Table 1 Examples of Tibetan La case sentences

句型	实例
业格句	ནགས་ཚལ་དུ་ཐང་ཤིང་གཅོད། 译：在森林里砍树。
为格句	བོད་ཡུལ་གྱི་འཚེལ་རྒྱས་ལ་རོགས་སྐྱོར་མང་པོ་བྱས་ཡོད། 译：为藏区的发展给予了很多帮助。
依格句	དཔལ་ནགས་ཚལ་དུ་གཙོས་ཤིང་མང་པོ་སྐྱེས་ཡོད། 译：今年在森林里长着很多松树。
同格句	རྒྱ་ཡིག་གི་གསར་འགྱུར་རྣམས་བོད་ཡིག་ཏུ་བསྒྱུར། 译：把汉文新闻翻译成了藏文。
时格句	རྒྱལ་ཡོངས་ཀྱི་ཚོགས་འདུའི་སྐབས་སུ་གཏམ་བཤད་གནང་། 译：在开国全会议时进行演讲。

在《藏语语法疑难释义》和《藏语语法研究》中提到不同 La 格虚词在各类用法（La 格句子）中的语义功能和使用度都有差异[1,17]，故本文对其进行了总结和分析：

（1）La 格中自由虚词不一定可以随机替换不自由虚词

在 La 格句子中，根据 La 格虚词的添接规则，原则上自由虚词“ན”或“ལ”可以随机替换其余 5 个不自由虚词“སྤ་ར་ན་ཏུ་ད”，但在实际使用中会出现不可替换或替换后不恰当的现象。如“གཟུགས་སུ་བཞུས།”替换成“གཟུགས་ལ་བཞུས།”后句型会发生改变，“ལུས་ལ་གོས་སྐོན།”替换成“ལུས་སུ་གོས་སྐོན།”后会出现 La 格虚词使用不恰当的现象。

（2）La 格中自由虚词的使用形自由义不自由

在 La 格虚词的不同用法中，自由虚词“ན”和“ལ”的使用虽在形式上自由，但在语义上不自由。如“སྒེར་ལ་བཞག”替换成“སྒེར་ན་བཞག”后会出现 La 格虚词使用不恰当的现象，“དགོན་པ་ན་སློབ་སྦྱོང་ཞིང།”替换成“དགོན་པ་ལ་སློབ་སྦྱོང་ཞིང།”后语义会发生变化。

（3）La 格虚词在不同用法中的使用度不同

相比其它 La 格虚词，自由虚词“ན”和“ལ”分别在依格和业格中的使用度较高，而在同格和时格中的使用度则偏低[1,17]。另外不自由虚词“སྤ་ར་ན་ཏུ་ད”在不同 La 格句子中的使用度也有所不同[1,17]。

1.2 La 虚词语义信息标记算法

La 格根据虚词的语义功能和添接规则分成了不自由虚词和自由虚词两种，其中，“སྤ་ར་ན་ཏུ་ད”为不自由虚词，添接受前一音节后加字的限制，需在后加字为“ས”的音节后添接“ས”；后加字为“ག་ད”和再后加字“~ད”之一的音节后添接“ཏ”；后加字为“ང་ད་ན་མ་ར་ལ”之一的音节后添接“ད”；后加字是“འ”或没有后加字的音节后添接“ར”或“ན”。“ན”和“ལ”是自由虚词，添接不受前一音节后加字的限制，可自由添接。La 格虚词的详细添接规则见表 2.

表 2 藏文 La 格虚词的添接规则

Table 2 Adding rules of The Tibetan La case function words

后加字	La 格虚词	
	不自由虚词	自由虚词
ས	ས	ན 或 ལ
ག、ད 或~ད	ཏ	
ང、ད、ན、མ、ར 或 ལ	ད	
འ 或无	ར 或 ན	

虽然所有 La 格虚词的总体用法一致，但因藏文 La 格虚词的语义功能和语言表达能力强于其它藏文虚词，在具体用法中的语义功能和使用度均会有一定的差异[1]。故在藏文 La 格分类任务的建模中，若模型除了获取每个 la 格句子的文本表示外，还能获取不同 La 格虚词的语义信息，将有利于提升最终 La 格句子的分类性能。基于此，本文将通过设计 La 格虚词语义信息标记算法，标记 La 格句子中 La 格虚词和其它音节的语义信息，以供模型在训练时学习到更加丰富和多样的语义特征。藏文 La 格虚词语义信息标记算法的主要功能是根据藏文 La 格虚词的用法及添接规则，在识别出所有输入 La 格句子中 La 格虚词的基础上，标记每个 La 格虚词的语义信息，并为了便于后续神经网络的建模，为其余音节都标记了语义信息“O”。藏文 La 格虚词语义信息标记算法详见算法 1.

算法 1 藏文 La 格虚词语义信息标记算法：

Input: S_D，L #S_D 是至少含一个 La 格虚词的 La 格句子集，L ← [‘ས’,‘ར’,‘ན’,‘ཏ’,‘ད’,‘ལ’,‘འ’]
Output: S_T #S_T 为标记好 La 格虚词和其它音节语义信息的 La 格句子集

```
S_T = []
for S in S_D do
    N_S = L_Tagger(S,L) #调用 L_Tagger 将已标记 La 格虚词语义信息的句子付给 N_S
    for C in N_S do #遍历句子中的每个音节
```

```
    if '/' in C then #判断'/'是否在音节中，若在说明是已标记语义的 La 格虚词
      N_S[N_S.index(C)] = (C.split('/')[0], C.split('/')[1]) #将'ཏུ/L4'等形式替换成('ཏུ','L4')的形式
    else #将 La 格虚词之外的音节都替换成('C','O')的形式
      N_S[N_S.index(C)] = N_S[N_S.index(C)] + ',' + 'O'
  S_T.append(N_S) #将标记好 La 格虚词和其它音节语义信息的句子逐条添加到列表 S_T 中
return S_T
function L_Tagger(Sentence,L) #定义 L_Tagger 函数用以标记每个句子中 La 格虚词的语义信息
  S ←Sentence.split()
  la_list = [Char for Char in S if Char in L] #获取每个句子中潜在的 La 格虚词
  if la_list.lenght >= 2 then #判断潜在的 La 格虚词个数是否大于等于 2
    for la in la_list do
      if la 的用法满足表 2 中不自由虚词的添接规则 then
        S[S.index(la)] = S[S.index(la)] + '/L' + str(La.index(la) + 1) #将 La 格虚词替换成'ཏུ/L4'的形式
      else #潜在 La 格虚词为自有虚词'ར'或'ལ'的情况
        S[S.index(la)] = S[S.index(la)] + '/L' + str(La.index(la) + 1)
    else #只有一个潜在 La 格虚词的情况
      S[S.index(la)] = S[S.index(la)] + '/L' + str(La.index(la) + 1)
```

算法 1 中的 S_D 表示已完成音节切分的藏文 La 格句子集，如“ཐོན་ཀྱང་ཐོག་མ་ཡིན།”等。S_T 是掉用函数 *La_ Tagger* 标记好 La 格虚词和其它音节语义信息后返回的 La 格句子集，如“[('ཐོན', 'O'), ('ཀྱང', 'O'), ('ཏུ','L7'), ('ཐོག', 'O'), ('མ', 'O'), ('ཡིན', 'O'), ('།', 'O')]”等。“L1，L2，…，L7”依次表示 La 格虚词“ཀུ་ར་རུ་ཏུ་ན་པ་ཏུ”的类别语义信息，“O”表示其他音节的语义信息。

2 融合 La 格虚词语义信息的藏文 La 格分类模型

本文模型的设计思路是在输入音节序列特征的基础上，额外加入 La 格虚（用 L_i 标记七个 La 格虚词中的第 i ∈[1,7]个 La 格虚词的语义信息）和其它音节（用“O”标记其它音节）的语义信息来增强输入部分的语义表达，进而达到优化模型性能的效果。基于此，我们提出了一种融合 La 格虚词语义信息的藏文 La 格分类模型，总体模型架构如图 1 所示。主要由六部分组成，分别如下：

（1）输入层：输入利用算法 1 标记好 La 格虚词和其它音节语义信息的藏文 La 格句子，输入单元为音节；

（2）嵌入层：将每个音节及对应 La 格虚词或其它音节的语义信息“L_i”或“O”映射成低维语义向量；

（3）卷积层：为了避免破坏特征的时序信息，模型仅采用一维卷积，通过拼接嵌入层的每个音节及对应 La 格虚词或其它音节的语义特征向量，完成每个音节及对应 La 格虚词或其它音节语义信息的融合，进而提取输入文本的空间语义特征；

（4）双向 LSTM 层：以提高时序特征的学习能力为目的，本文在卷积操作后直接拼接双向 LSTM 学习文本时序特征；

（5）注意力机制层：为了充分利用双向 LSTM 层每个时刻的输出，使用注意力机制（Attention）对双向 LSTM 层各时刻的输出特征进行加权融合；

（6）分类层：将注意力机制层输出的句子级语义向量输入到全链接层和 Softmax 层进行最终的 La 格分类。

2.1 嵌入层

给定一个包含 T 个音节及对应 La 格虚词或其它音节语义信息标记的 La 格句子 $S=\{(c_1,t_1),(c_2,t_2),\cdots,(c_n,t_T)\}$，为了将 S 中的每个音节 c_i 及对应 La 格虚词或其它音节的语义信息 t_i 映射成实值向量 e_i^c 和 e_i^t，需要分别从音节嵌入矩阵 $W^{char}\in\mathbb{R}^{d^c\times|V|}$ 及对应 La 格虚词或其它音节语义信息的语义特征嵌入矩阵 $W^{tag}\in\mathbb{R}^{d^t\times|V|}$ 中查找 S 中的每个 c_i 及 t_i，其中矩阵 W^{char} 和 W^{tag} 是模型要学习的参数，V 是词汇表大小，d^c 是音节嵌入的大小，d^t 是音节语义特征嵌入的大小，与 d^c 的大小相等。所以我们可以使用矩阵和向量的乘积将 c_i 和 t_i 映射成 e_i^c 和 e_i^t：

$$e_i^c = W^{char}v^i \quad (1)$$

$$e_i^t = W^{tag}v^i \quad (2)$$

其中，v^i 是大小为|V|的向量，在 e_i^c 和 e_i^t 处的索引值为 1，其他位置的值均为 0。至此，藏文 La 格句子将可以作为实值向量 $emb_s=\{(e_1^c,e_1^t),(e_2^c,e_2^t),\cdots,(e_T^c,e_T^t)\}$ 送入模型。

2.2 卷积层

为了增加文本表示的空间维度和特征多样性，达到丰富特征表达的目的，本节将音节嵌入 e_i^c 及对应 La 格虚词或其它音节的语义特征嵌入 e_i^t 进行拼接后作为卷积层的输入，采用一维卷积提取固定感受视野下的局部空间特征，并完成对 e_i^c 和 e_i^t 两种语义信息的融合，融合语义信息的形式化表示如下：

$$g_0^f,\cdots,g_T^f=\mathrm{CONV}_k([e_1^c,e_1^t],\cdots,[e_T^c,e_T^t]) \quad (3)$$

其中，CONV_k 表示一维卷积层，k 是卷积核大小，即感受视野。

2.3 双向 LSTM 层

在句子和短文本分类任务中，卷积之后直接使用池化操作容易造成文本时序信息的损失，进而影响模型性能[14,25]。为了避免卷积之后直接进行池化操作而对时序特征造成破坏，将在卷积操作后拼接双向 LSTM 来学习文本的时序特征，以提高模型对上下文时序信息的学习能力。

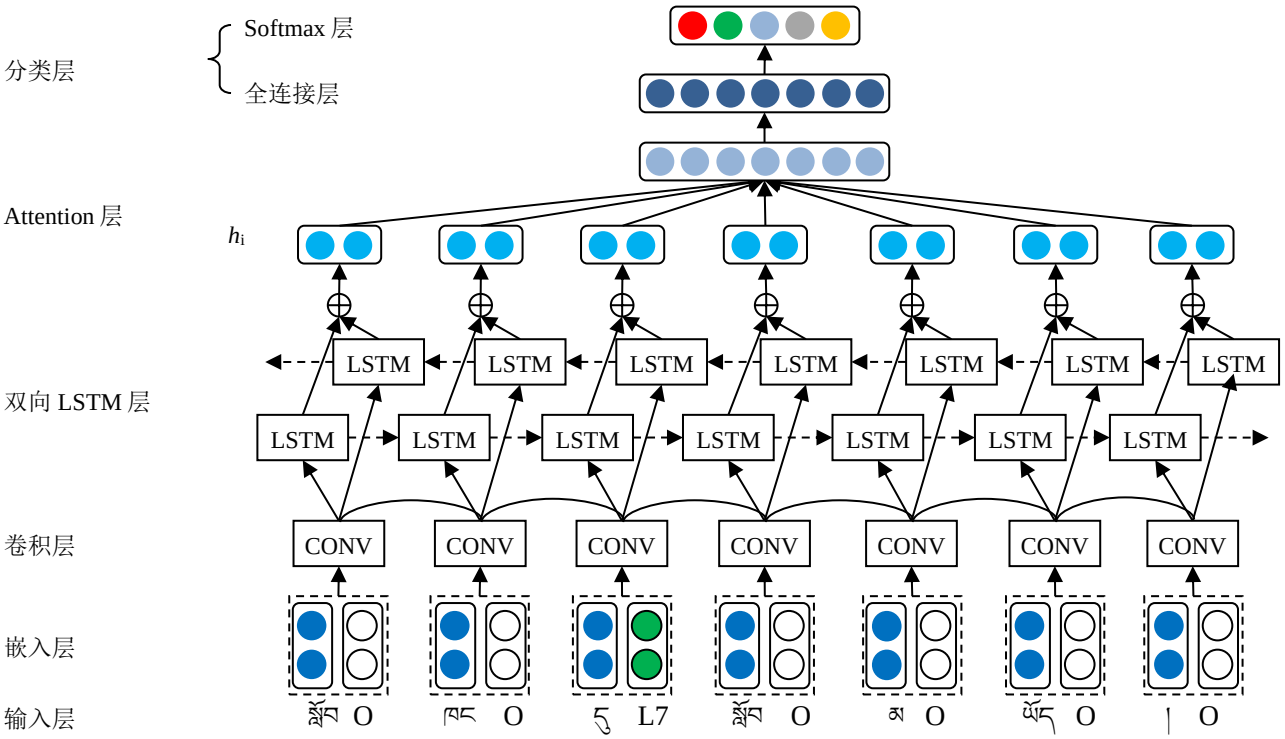

图 1 融合 La 格虚词语义信息的藏文 La 格分类模型架构

Fig.1 Tibetan La case classification model architecture with Fused La Case Function Word Type Information

双向 LSTM 层的输入向量为卷积层的输出特征，所以 t 时刻双向 LSTM 层的最终输出可以表示为 h_t，其计算过程如下：

$$\overrightarrow{h_1},\cdots,\overrightarrow{h_T}=\mathrm{LSTM}_f(g_1^f,\cdots,g_T^f) \quad (4)$$

$$\overleftarrow{h_1},\cdots,\overleftarrow{h_T}=\mathrm{LSTM}_b(g_1^f,\cdots,g_T^f) \quad (5)$$

$$h_t=\left[\overrightarrow{h_t}+\overleftarrow{h_t}\right] \quad (6)$$

其中，LSTM_f 和 LSTM_b 分别表示正向 LSTM 和反向 LSTM，$\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$ 分别表示 t 时刻 LSTM 的正向和反向的输出值，$t\in[1,T]$。

2.4 注意力机制层

为了充分将注意力聚焦在 LSTM 层各时刻输出特征中贡献较大的文本特征上，本文在双向 LSTM 层后采用注意力机制对各个时刻的输出特征进行加权融合。

假设双向 LSTM 层 t 时刻的输出向量用 H_t 表示，e_t 表示 H_t 对 La 格句子语义的重要程度，a_t 表示 H_t 对整个 La 格句子语义表示贡献的权重。根据上述定义，可得注意力权重的计算方法如式（7）和式（8）所示。

$$e_i=u^T\cdot\tanh(W_a\cdot H_t+b_a) \quad (7)$$

$$a_t=\frac{\exp(e_t)}{\sum_{j=0}^{T}\exp(e_j)} \quad (8)$$

其中，u^T、W_a 和 b_a 是模型需要学习的参数，tanh 为非线性激活函数。

通过式（8）能够计算出双向 LSTM 层所有时刻的注意力权重，然后对其进行加权求和便可得到注意力机制层最终输出的特征向量 V：

$$V=\sum_{t=0}^{T}a_t\cdot H_t \quad (9)$$

其中，$T=\{t_0,\ t_1,\ \cdots t_{n-1}\}$，n 是每条藏文 La 格句子包含的音节数，即句子长度。

2.5 分类层

得到注意力机制层的最终输出 V 后，就可以通过全连接层和 Softmax 层预测 La 格的类型。用 $\boldsymbol{S}(\boldsymbol{S}\in S_i)$表示 V 经过全连接层输出的文本语义特征向量，则模型将某条藏文 La 格句子预测为 i 类的概率为 P_i，其计算过程如下：

$$\mathbf{S}=softmax(W_i\cdot V+b_i) \quad (10)$$

$$p_i=\frac{\exp(s_i)}{\sum_{m=1}^{n}\exp(s_m)},n=5 \quad (11)$$

其中，$i\in[1,5]$，分别表示五种藏文 La 格句子类型，n 表示句型数目，S_i 表示 $\boldsymbol{S}$ 中属于类别 i 的分值，W_i 与 b_i 为全连接层需要学习的参数。

3 实验

3.1 实验环境与数据说明

3.1.1 实验环境

实验设计平台是 Anaconda，采用的框架为 Keras，后端为 Tensorflow 框架。其它配置参数见表 3.

表 3 实验环境配置

Table 3 Experimental environment configuration

实验环境	配置参数
操作系统	Ubuntu20.04.2LTS
开发语言	Python3.0
CPU/内存	Intel(R) Xeon(R) Gold 5112 @ 3.60GHz/128 GB
GPU/显存	NVIDIA Quadro P6000/24 GB
运行环境	CUDA9.0 CUDNN7.6.5

3.1.2 实验数据说明

为了确保实验结果的可对比性，所用实验数据是文献[14]中所构建的 La 格分类数据集，我们以后续方便使用为目的，简称藏文 la 格分类数据集为 TLD。该数据集共有 20000 条 La 格句子，每条句子有且仅含一个 La 格虚词，其中业格句有 6964 条，为格句有 2684 条，依格句有 3104 条，同格句有 3595 条，时格句有 3653 条，分别占总数据集的 34.82%、13.42%、15.52%、17.98%和 18.26%。实验时，按 8:1:1 的比例将数据集 TLD 分成了训练集、验证集和测试集。

3.2 基线方法与参数设置

3.2.1 基线方法选择

目前仅有一篇有关藏文 La 格自动分类的文献报道，若只选择该文献为基线验证本文模型的有效性，则会显得基线方法偏少，致使实验的说服力大打折扣。所以为了充分验证本文模型的效果，我们选择了两类的基线方法，为了便于下文写作，我们称第一类为基线一，第二类为基线二。基线一是在句子和短文分类任务中常用的7个经典基线模型，基线二是文献[14]中的模型，是仅有的一篇有关藏文 La 格自动分类的文献报道。

（1）FastText：是 Facebook 于 2016 年提出的一种快速文本分类工具，出自文献[18]。

（2）TextRNN：是一种运用于多标签分类问题的方法，结构非常灵活，出自文献[19]。

（3）Bi-LSTM：是一种采用双向 LSTM 进行关系分类的方法，在文本分类任务中取得了理想的性能，出自文献[20]。

（4）Bi-LSTM+Attention：是一种基于注意力机制的双向长短时记忆关系分类网络，在文本分类任务中也取得了理想的性能，出自文献[21]。

（5）TextCNN：是应用于句子分类任务的首个卷积神经网络模型，为句子级分类任务提供了便利，出自文献[22]。

（6）TextRCNN：是一种应用于文本分类任务的循环卷积神经网络，集成了 RNN 和 CNN 的优点。出自文献[23]。

（7）C-LSTM：是一种使用单通道的多路卷积加双向 LSTM 进行文本分类的方法，出自文献[24]。

（8）SF-C+LSTM+Att：是一种融合双通道音节特征的单向 LSTM 藏文 La 格分类模型，出自文献[14]。

（9）SF-C+Bi-LSTM+Att：是一种融合双通道音节特征的双向 LSTM 藏文 La 格分类模型，出自文献[14]。

FWS-C-Bi-LSTM+Att是本文模型，FWS表示融合La格虚词语义信息，C 表示一维卷积操作，Att 表示注意力机制（Attention）。

3.2.2 实验参数设置

在实验过程中，为了确保实验结果的可对比性，对所有模型的超参数进行了调参范围限定[25]，经过多次调参，最终在有限的范围内选择了当前最优的超参数组合，本文模型的主要参数见表 4.

表 4 模型参数

Table 4 Model parameters

参数名	参数	参数名	参数
最长音节数	20	损失函数	categorical_crossentropy
批处理大小	16	辍学率	0.25
卷积核数量	200	学习率	0.0001
卷积核大小	3	优化函数	Adam
LSTM 大小	128	迭代次数	40

3.3 实验结果与分析

3.3.1 各模型的分类性能对比

为了验证本文方法的有效性和优越性，分别在两类基线方法上对比了藏文 La 格的分类效果，选用的评价指标分别是精度(*P*)、召回率(*R*)、F_1 值和准确率(ACC)，实验结果见表 4.

表 4 藏文 La 格分类实验结果（%）

Table 4 Experimental results of Tibetan La case classification (%)

实验	模型	*P*	*R*	F_1	ACC
基线一	FastText	88.67	87.51	88.09	88.10
	TextRNN	87.73	86.56	87.14	87.16
	Bi-LSTM	88.07	86.54	87.45	87.50
	Bi-LSTM+Att	88.26	86.93	87.59	87.75
	TextCNN	89.18	87.68	88.42	88.50
	TextRCNN	86.08	84.73	85.40	85.50
	C-LSTM	88.83	87.76	88.29	88.34
基线二	SF-C+LSTM+Att	88.83	87.76	88.29	88.34
	SF-C+Bi-LSTM+Att	90.58	89.77	90.17	90.26
本文	FWS-C-Bi-LSTM+Att	**93.14**	**93.01**	**93.07**	**93.10**

从表 4 中可以看出，相较基线一中的七种方法，本文模型的分类准确率平均提高了 5.55%，相较基线二中的两种方法，本文模型的分类准确率平均提高了 3.8%，表明本文模型取得了更好的分类效果，验证了本文模型的优越性。原因有四，一是本文使用藏文 La 格虚词语义信息标记算法，增加了输入特征的多样性，丰富了嵌入向量的语义信息；二是采用一维卷积操作对每个音节及对应 La 格虚词或其它音节的语义特征向量进行了融合，增加了文本表示的空间维度，提高了文本表示的空间特征质量；三是采用双向 LSTM 学习时序特征，更合理地学习了文本时序特征，提高了文本的时序特征质量；四是采用注意力机制计算双向 LSTM 层每一时刻输出特征的贡献值，并进行加权融合，更加充分地学习了每一时刻的输出特征，提高了最终文本表示的质量。

本文经分析实验结果发现，主要影响模型性能欠佳的原因有二，一是部分 La 格句子无法仅凭上下文时序特征、语法结构和浅层语义信息进行分类，如：“གཟུགས་ལ་བཞེས”和“ས་ལ་བསྒེལ་ནས་བཞག”属于业格，而“གཟུགས་སུ་བཞེས”和“ས་ལ་འགྲེལ་ནས་འདུག”分别属于同格和依格。“བ་ཡང་ལ་དཀར་ཟལ་ཡོད།”和“བ་ཡང་ལ་དཀར་ཐེག་ཡོད།”中，前一句属于同格，而后一句属于依格。可见，上述句子需要根据具体的语境、语用目的和深层语义来判断其类别。二是有些藏文La格句子存在兼类现象，如：“ཐོན་དུ་འགོ།”、“གཞུག་ཏུ་ཐོས།”和“ཕྱི་རྒྱལ་འཛུང་།”等 La 格句子可以根据对“ཐོན”、“གཞུག”和“ཕྱི”的不同理解分为业格或时格。可见，类似于上述藏文 La 格句子需要理解其深层语义信息和具体的语用目的才能准确分类。

3.3.2 特征学习方式对模型性能的影响

为了验证本文方法中融合 La 格虚词语义信息方法的有效性以及模型拼接方式的科学性，一是比较了使用 La 格虚词语义信息标记算法前后模型的分类性能，实验结果见图 2(a).二是比较了不同卷积模式对模型分类性能的影响，实验结果见图 2(b).三是比较了使用单向 LSTM 时

和双向 LSTM 时模型的分类效果，实验结果见图 2(c).四是比较了加或不加注意力机制时模型的分类效果，实验结果见图 2(d).

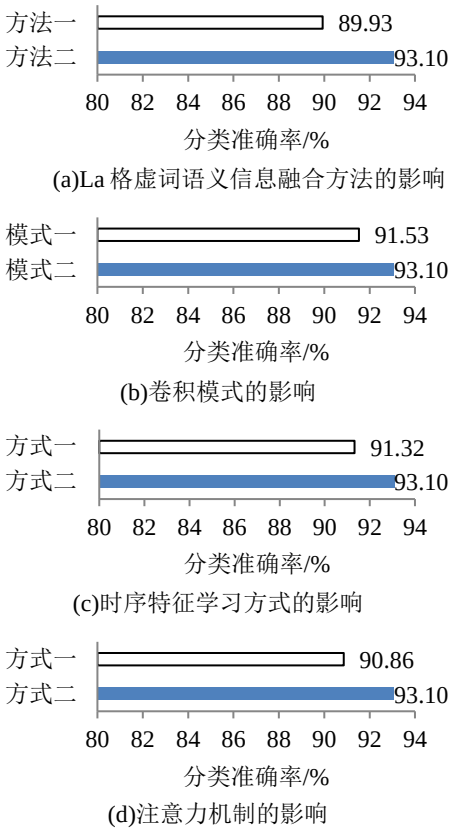

图 2 特征学习方式对模型性能的影响

Fig.2 Influence of feature learning methods on model performance

图 2(a)中的方法一和方法二分别表示使用 La 虚词语义信息标记算法前后模型的分类效果。可以看出，使用 La 格虚词语义信息标记算法后的藏文 La 格分类准确率比使用前高 3.17%，验证了 La 格虚词语义信息标记算法的有效性；图 2(b)中的模式一和模式二分别表示使用标准一维卷积神经网络时和只使用一维卷积操作时模型的分类效果。可以看出，只使用一维卷积操作时模型的藏文 La 格分类准确率比使用标准一维卷积神经网络时高 1.57%，表明不使用池化法的一维卷积模式更有利于提升模型性能；图 2(c)中的方式一和方式二分别表示卷积后直接拼接单向和双向 LSTM 时模型的分类效果，可以看出，卷积后直接拼接双向 LSTM 时的藏文 La 格分类准确率比卷积后直接拼接单向 LSTM 时高 1.78%，表明卷积后拼接双向 LSTM 学习时序特征的方法更有效；图 2(d)中方式一和方式二分别表示在双向 LSTM 后不加或加注意力机制时模型的分类效果。可以看出，加注意力机制时的分类准确率比不加时高 2.24%，表明本文模型在双向 LSTM 后加注意力机制时，可以充分利用双向 LSTM 层每一时刻的输出特征，进而提高模型最终获取的特征质量，达到提高模型分类性能的效果。

3.3.3 融合 La 格虚词语义信息前后的对比实验

为了验证本文设计的 La 格虚词语义信息标记算法的通用性和易用性，分别在基线一的 7 种分类模型上对比了融合 La 格虚词语义信息前后的效果，方式一为没有融合 La 格虚词语义信息时的分类准确率，方式二为融合 La 格虚词语义信息后的分分类准确率，结果详见图 3.

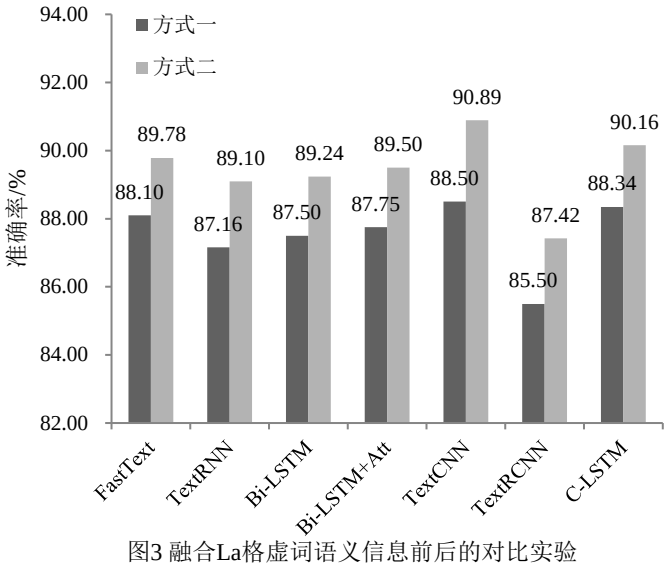

图3 融合La格虚词语义信息前后的对比实验

Fig.3 A comparative experiment before and after fusing the type information of La case function words

从图 3 中的实验结果可以看出，使用藏文 La 格虚词语义信息标记算法后，各种神经网络分类模型的藏文 La 格句子分类效果均优于使用前的效果，分类准确率平均提高了 1.89%，验证了藏文 La 格虚词语义信息标记算法的通用性和易用性。另外，因 CNN 网络擅长捕获局部特征信息，而藏文 La 格句子的类型特征往往会以 La 格虚词为中心集中在句子的某个局部，故 CNN 的藏文 La 格句子分类性能较高。

4 结论

基于藏文 La 格句子分类任务在藏语自然语言处理领域的重要性，本文通过分析 La 格虚词的语义特征及用法，在设计 La 格虚词语义信息标记算法的基础上，提出了一种融合 La 格虚词语义信息的藏文 La 格分类模型，该模型首先通过设计 La 格虚词语义信息标记算法，为模型融合每个音节及对应 La 格虚词或其它音节的语义信息奠定了基础；然后通过汲取 CNN、LSTM 和 Attention 提取文本特征的优点，进行神经网络模型的拼接，完成了每个音节与对应 La 格虚词或其它音节语义信息的融合和整个模型的训练；最后经过设计三组实验，一是在 TLD 数据集上与基线模型进行比较，验证了本文模型的优越性；二是通过考查融合每个音节及对应 La 虚词或其它音节语义信息前后和不同特征学习方式对模型性能的影响，验证了融合 La 格虚词和其它音节语义信息的有效性及本文模型拼接方式的科学性；三是经过在七种常见分类神经网络模型上对比融合 La 格虚词语义信息前后 La 格的分类效果，验证了藏文 La 格虚词语义信息标记算法的通用性和易用性。

未来，将通过进一步扩充 TLD 数据集，研究输入以词为基元的藏文 La 格分类方法，并尝试融入每个词的语义特征信息及更深层次的句法信息等，以进一步优化模型的分类性能。

参考文献

[1] 吉太加.藏语语法疑难释义[M].民族出版社,北京,2017.
[2] 万齐斌,董方敏,孙水发.基于 BiLSTM-Attention-CNN 混合神经网络的文本分类方法[J]. 计算机应用与软件,2020,37(09):94-98+201.
[3] 梁顺攀,豆明明,于洪涛,等.基于混合神经网络的文本分类方法[J].计算机工程与设计,2022,43(02):573-579.
[4] Kim Y. Convolutional Neural Networks for Sentence Classification[C]. // Proceedings of Empirical Methods on Natural Language Processing. Doha, 2014: 1746-1751.
[5] Zhao Z, Wu Y. Attention-Based Convolutional Neural Networks for Sentence Classification[C]. //Interspeech. San Francisco, 2016: 705-709.
[6] Vieira J, Moura R S. An analysis of convolutional neural networks for sentence classification[C]. // Computer Conference. IEEE, 2017.
[7] 高云龙,左万利,王英,等.基于稀疏自学习卷积神经网络的句子分类模型[J].计算机研究与发展,2018,55(001):179-187.
[8] Yang L, Ji L, Huang R, et al. Multi-Grained-Attention Gated Convolutional Neural Networks for Sentence Classification[J]. Intelligent Data Analysis, 2019, 23(5): 1091-1107.
[9] 汪冉,金忠.基于极性转移和 LSTM 的树结构网络与句子分类[J].计算机应用研究,2019,36(01):64-67.
[10] Fentaw H W, Kim T H. Design and Investigation of Capsule Networks for Sentence Classification[J]. Applied Sciences, 2019, 9(11): 2200.
[11] 李文宽,刘培玉,朱振方,等.基于卷积神经网络和贝叶斯分类器的句子分类模型[J]. 计算机应用研究,2020,37(02):333-336+341.
[12] 柔特,才让加.基于循环卷积神经网络的藏文句类识别[J].中文信息学报,2019,33(12):76-82.
[13] Ban M, Cai Z, Cai R, et al. Tibetan interrogative sentence recognition and classification based on phrase features[J]. MATEC Web of Conferences, 2021, 336(4): 06017.
[14] 班玛宝,才让加,张瑞,等.融合双通道音节特征的藏文 La 格例句自动分类模型[J]. 北京大学学报(自然科学版),2022,58(01):91-98.
[15] 凌逆战.CNN 神经网络之一维卷积、二维卷积详解[BD/OL].https://www.freesion.com/article/17891324479/
[16] 马世拓,班一杰,戴陈至力.卷积神经网络综述[J].现代信息科技,2021,5(02):11-15.
[17] 吉太加.藏语语法研究[M].青海民族出版社,青海,2016.
[18] 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). Valencia, 2017: 427–431.
[19] LIU P, QIU X, HUANG X. Recurrent Neural Network for Text Classification with Multi-Task Learning[C]. // Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016: 2873–2879.
[20] Zhang S, Zheng D, Hu X, et al. Bidirectional Long Short-Term Memory Networks for Relation Classification [C]. // Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. Shanghai, China, 2015: 73-78.
[21] Peng Z, Wei S, 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 (Volume 2: Short Papers). Berlin, Germany, 2016: 207-212.
[22] Kim Y. Convolutional neural networks for sentence classification[C]. // Proceedings of Empirical Methods on Natural Language Processing. Doha, Qatar, 2014: 1746–1751.
[23] Lai S, Xu L,Liu K, et al. Recurrent convolution neural networks for text classification[C]. // Proceedings of the 29th AAAI Conference on Atrificial Intelligence. Texas, USA, 2015: 2267-2273.
[24] Zhou C, Sun C, Liu Z, et al. A C-LSTM neural network for text classification[J]. Computer Science, 2015, 1(4): 39-44.
[25] 韩永鹏,陈彩,苏航,等.融合通道特征的混合神经网络文本分类模型[J]. 中文信息学报,2021,35(2):78–88.

Tibetan La Case Classification Model with Fused La Case Function Word Semantic Information

BAN Ma-bao, CI Zhengjiacuo[1,2], ZHANG Rui, CAI Rang-jia*[1,2,3,4,5]

(1.College of Computer Science and Technology, Qinghai Normal University, Qinghai Xining 810016; 2.The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Qinghai Xining 810008; 3.Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province, Qinghai Xining 810008; 4.Key Laboratory of Tibetan Information Processing, Ministry of Education, Qinghai Xining 810008; 5.Tibetan Information Processing Engineering Technology and Research Center of Qinghai Province, Qinghai Xining 810008)

Abstract: Using machine learning method to classify Tibetan La case (ལ་དོན།) is a challenging and important Tibetan natural language processing task. Based on the fact that the automatic classification of Tibetan La case is more dependent on the temporal nature of context semantic information and features, this paper by analyzes the semantic features and usage of La case function words, and on the basis of designing the semantic information labeling algorithm of La case function words, a Tibetan La case classification model with fused La case function words semantic information is proposed. The model first takes each syllable and the corresponding La case function word or other syllable semantic feature embedding as input, enriches the semantic information of the embedding vector, and increases the diversity of input features; Secondly, one-dimensional convolution fusion is used to learn the local feature vectors of each syllable and the corresponding La case function words or other syllable semantic information, so as to improve the spatial feature learning ability of convolution layer; Then, two-way LSTM is used to learn temporal features to improve the learning ability of temporal features; Finally, the attention mechanism is used to weight and fuse the output features of the two-way LSTM layer at each time, and make full use of the output features at each time to improve the feature quality of the final text representation. Experiments on TLD data sets show that the classification effect of the method is better than that of the baseline model and the model embedded only with syllables, the classification accuracy on the test set is 93.10 %.

Keywords: NLP; La case function word; Semantic information; neural network; La case classification