



一种集成深度学习模型的旅游问句文本分类算法

马喆康^{a,c}, 迪力亚尔·帕尔哈提^{a,c}, 早克热·卡德尔^{b,c}, 吐尔根·依布拉音^{b,c},
西尔艾力·色提^{b,c}, 艾山·吾买尔^{b,c}

(新疆大学 a. 软件学院; b. 信息科学与工程学院; c. 新疆多语种信息技术重点实验室, 乌鲁木齐 830046)

摘 要: 为提高旅游问句文本中关键特征的利用率, 提出一种集成词级卷积神经网络(WL-CNN)与句级双向长期记忆(SL-Bi-LSTM)网络的旅游问句文本分类算法。利用 WL-CNN 和 SL-Bi-LSTM 分别学习词序列子空间向量和句序列深层语义信息, 通过多头注意力机制将两种深度学习模型进行集成以实现旅游问句文本的语法和语义信息互补, 并通过 SoftMax 分类器得到最终的旅游问句文本分类结果。实验结果表明, 与基于传统深度学习模型的旅游问句文本分类算法相比, 该算法在准确率和损失率上分别取得了 0.986 6 和 0.127 7 的最优结果, 具有更好的分类效果。

关键词: 子空间结构信息; 深层语义信息; 多头注意力机制; 卷积神经网络; 双向长短期记忆网络

开放科学(资源服务)标志码(OSID):



中文引用格式: 马喆康, 迪力亚尔·帕尔哈提, 早克热·卡德尔, 等. 一种集成深度学习模型的旅游问句文本分类算法[J]. 计算机工程, 2020, 46(11): 70-76.

英文引用格式: MA Zhekang, Diliyaer Paerhati, Zaokere Kadeer, et al. A classification algorithm for tourist question texts integrated with deep learning models[J]. Computer Engineering, 2020, 46(11): 70-76.

A Classification Algorithm for Tourist Question Texts Integrated with Deep Learning Models

MA Zhekang^{a,c}, Diliyaer Paerhati^{a,c}, Zaokere Kadeer^{b,c}, Tuergen Yibulayin^{b,c}, Xerali Setti^{b,c}, Aishan Wumaier^{b,c}

(a. College of Software; b. College of Information Science and Engineering;

c. Xinjiang Key Laboratory of Multi-language Information Technology, Xinjiang University, Urumqi 830046, China)

[Abstract] To improve the utilization of key features of tourist question texts, this paper proposes a classification algorithm for tourist question texts integrated with the Word Level Convolutional Neural Network (WL-CNN) and the Sentence Level Bi-directional Long Short-Term Memory (SL-Bi-LSTM) network. The algorithm uses WL-CNN and SL-Bi-LSTM to learn the subspace vector of the word sequence and the deep semantic information of the sentence sequence. Then the two deep learning models are integrated by using the Multi-Head Attention Mechanism (MH-AM) to realize the syntactic and semantic information complementary of tourist question texts. Finally, the SoftMax classifier is used to obtain the classification results of tourist question texts. Experimental results show that the proposed algorithm has better classification performance than the tourist question text classification algorithms based on traditional deep learning models, increasing the accuracy to 0.986 6 and loss rate to 0.127 7.

[Key words] subspace structure information; deep semantic information; Multi-Head Attention Mechanism (MH-AM); Convolutional Neural Network (CNN); Bi-directional Long Short-Term Memory (Bi-LSTM) network

DOI: 10.19678/j.issn.1000-3428.0055990

基金项目: 国家自然科学基金(61762084); 国家重点研发计划(2017YFB1002103); 新疆维吾尔自治区重点实验室开放课题(2018D04019)。

作者简介: 马喆康(1995—), 男, 硕士研究生, 主研方向为自然语言处理、问句文本分类; 迪力亚尔·帕尔哈提, 硕士研究生; 早克热·卡德尔(通信作者), 实验师、硕士; 吐尔根·依布拉音, 教授、博士; 西尔艾力·色提, 硕士研究生; 艾山·吾买尔, 副教授、博士。

收稿日期: 2019-09-11 **修回日期:** 2019-11-06 **E-mail:** zuhra@xju.edu.cn

0 概述

随着我国社会经济的发展和人们物质生活水平的提高,旅游已经成为人们休闲娱乐的主要方式,但游客在旅游过程中发生的路线规划、酒店预订等问题也不断增加^[1]。目前,游客主要通过旅游网站问答方式获取旅游信息,需要对问题进行发布并等待其他用户的回复,具有延时性,而且旅游问答社区通常根据地理位置的问题分类方式,无法全面覆盖问题的所有类别。此外,传统旅游问答社区一般采用人工标注或机器学习模型进行问题分类,导致分类效率和准确率均较低,无法快速精准地定位游客的问题类别,进而影响后续的信息检索。因此,如何快速高效地对各类旅游问句进行自动分类已成为亟待解决的问题。深度学习技术在近几年得到快速发展,且被应用于问句文本分类任务中并取得了较好的成果^[2],相比传统机器学习技术更能捕获文本的深层语义信息及解决人工设计特征导致的误差问题,且分类精度更高^[3],但其多数为基于单一结构的深度学习模型或仅对多个模型进行简单串联,因此在挖掘文本深层特征时,会丢失大量的语法和句法信息并增加冗余信息。

本文提出一种集成多种深度学习模型的旅游问句文本分类算法,通过词级卷积神经网络(Word Level Convolutional Neural Network, WL-CNN)获取词的低层空间结构信息构建文本的语法信息,利用句级双向长短期记忆(Sentence Level Bi-directional Long Short-Term Memory, SL-Bi-LSTM)网络对旅游问句文本的全局语义和句法信息进行建模,同时结合多头注意力机制(Multi-Head Attention Mechanism, MH-AM)对这两种深度学习模型进行联合学习并分配注意力权重,以提高旅游问句文本关键特征的利用率。

1 相关工作

早期的问句文本分类方法主要利用简单的机器学习模型或深度学习模型对不同类型的问句文本进行分类识别。文献[4]提出一种体育文本自动分类的半监督机器学习模型并取得了87%的分类精度。文献[5]提出一种基于时间加权函数(Temporal Weighting Function, TWF)的文本自动分类方法,实验结果表明,与基于传统支持向量机(Support Vector Machine, SVM)的文本自动分类方法相比,该方法的分类准确度提高了17%。文献[6]提出一种向量空间模型,通过对阿拉伯语言文本中的问题进行形式化及特征约束等处理,实现阿拉伯语言文本的分类及主题匹配。文献[7]结合人工标注和多类SVM对不同文本进行筛选,设计一种主动学习分类模型实现大型高维文本的精确分类,并减少了人工标注的工作量。文献[8]融合词向量和BTM模型对问题文

本主题进行扩展,利用SVM进行问句分类。由于简单的机器学习模型难以捕获问句文本的深层抽象特征且受人工设计特征的误差影响,因此使其不能有效识别和处理相对复杂的问句文本。

深度学习技术的不断发展为问句文本分类提供了一种新的思路。文献[9]提出一种多层级注意力卷积长短期记忆(Multi-level Attention Convolutional Long Short-Term Memory, MAC-LSTM)网络的问句文本分类模型,实现问句文本的准确分类。文献[10]提出一种双向长短期记忆(Bi-directional Long Short-Term Memory, Bi-LSTM)网络的新闻文本分类识别方法,解决了新闻文本数据稀疏、维度爆炸以及人工设计特征参与标注带来的局限性等问题,提高了分类精度。文献[11]引入词向量模型并结合卷积神经网络(Convolutional Neural Network, CNN)进行问题分类,提高了未标注样本的利用率。文献[12]结合朴素贝叶斯支持向量机(Naive Bayes Support Vector Machine, NB-SVM)、词向量和长短期记忆循环神经网络(Recurrent Neural Network, RNN),提出一种混合深度学习模型的双语文本分类方法,提高了双语情绪文本的分类识别效果。文献[13]结合Word2vec和决策树并引入深度学习技术,实现了影评情感文本的准确分类。文献[14]针对邻域问题文本进行分类,并取得了较好的分类结果。

2 集成深度学习模型的旅游问句文本分类

旅游问句文本的句法和语义信息主要取决于文字组成和序列顺序。一方面,旅游问句的语法由多个疑问关键词和某些网络流行词组成,对文本序列中的词进行建模,构成文本序列的低层子空间结构信息。另一方面,旅游问句文本的语义信息和句法信息来自文本序列本身,因此本文使用one-hot方式对旅游问句文本进行编码建模,捕获文本的语义信息和句法信息。

为更好地表示旅游问句文本,本文提出一种集成深度学习模型的旅游问句文本分类算法。该算法主要由词级卷积神经网络、句级双向长短期记忆网络和多头注意力机制三部分组成,通过两种深度学习模型获取旅游问句文本的局部深层语义特征和低层子空间结构信息以及全局语义信息和深层结构信息,并利用多头注意力机制对其进行合并分类,实现两种深度学习模型在语言的句法和语义方面保持互补关系。

2.1 词级卷积神经网络

目前,卷积神经网络在场景分类、模式识别和图像分割等领域均具有广泛应用^[15]。深度卷积神经网络中的卷积层和池化层不仅对文本中所含噪声具有很强的鲁棒性,而且其因为参数共享和局部连接特性可对文本的频谱或者语义进行建模^[16],所以本

文利用卷积神经网络对旅游问句文本中的词进行建模,进一步挖掘旅游问句文本中词的子空间结构信息。

在使用卷积神经网络提取旅游问句文本中词的相关特征时,将卷积层中预定义的 1-of- n 编码数字替换为旅游问句文本中的词^[17],且最大编码数字为 225。当使用文本中的词替换编码数字时,若每句文本中所含有的词不足以替换所有数字,则用数字 0 进行填充,即卷积神经网络中输入层的输入向量为 (Nan, 225, 225),其中 Nan 表示空值,然后输入初始卷积层中对旅游问句文本的词进行编码,具体步骤如下:

1) 假设编码后的输入向量为 x_i ,总输出向量为 z_{xy}^l , l 表示卷积神经网络的隐含层层数,卷积层的输出向量为 y_i^l ,滤波器数量为 k ,权重向量为 w 。若词 W 通过初始卷积层,则需满足:

$$v = \text{conv2}(w, x, \text{"padding"}) + b \quad (1)$$

其中, v 表示初始卷积层的输出, w 表示权重矩阵, b 表示偏置向量,padding 表示填充,其输出表示为:

$$y = \eta(v) \quad (2)$$

其中, $\eta(\cdot)$ 表示激活函数。

2) 通过多层卷积层可计算得到旅游问句文本的卷积词汇特征向量:

$$z_{xy}^l = \sum_i \sum_j w_{ij} y_{(x+i)(y+j)}^l \quad (3)$$

3) 通过最大池化层操作后可获得旅游问句文本中所有词的 CNN 特征向量:

$$p_{xy}^l = \max z_{xy}^{l-1} = \max \left(\sum_i \sum_j w_{ij} y_{(x+i)(y+j)}^l \right) \quad (4)$$

其中, p_{xy}^l 表示最大池化层的输出特征向量。

词级卷积神经网络在提取旅游问句文本中的词特征时,使用 5 层卷积和池化层,且每层神经元个数为 32、64、128、256 和 512,且卷积核和池化核的大小分别为 3×3 和 2×2 。为进一步提高词级卷积神经网络对文本中词的建模能力,在 5 层卷积池化层后连接 2 层全连接层,即全连层中的神经元个数为 512 和 64,从而实现旅游问句文本中词级子空间结构信息的提取。

由于词级卷积神经网络主要是用于捕获旅游问句文本中词的低层子空间结构信息和局部深层空间特征,因此需要对输入的旅游问句文本进行相关预处理,提高本文算法所提取特征的表征能力和抗噪能力。

2.2 句级双向长短期记忆网络

词级卷积神经网络主要获取旅游问句文本的词级低层子空间结构信息和局部深层语义信息,但忽略了旅游问句文本的全局语义特征,利用双向长短期记忆网络对旅游问句文本序列进行时空建模,

具体步骤如下:

1) 通过词嵌入技术^[18]将旅游问句文本表示为一个 one-hot 编码向量,并利用前向和后向长短期记忆层对旅游问句文本编码向量序列中的时间特征进行建模。

2) 利用双向长短期记忆(Long-Short Term Memory, LSTM)网络^[19-20]对输入序列 x 之间的关系进行映射,并通过不同时刻的神经元激活状态计算得到输出序列 y ,计算公式为:

$$i_t = \sigma(w_{ix}x_t + w_{ia}\alpha_{t-1} + w_{ic}c_{t-1} + b_i) \quad (5)$$

文本序列在 t 时刻通过输入门 i_t 后,需要遗忘不必要的文本信息,计算公式为:

$$f_t = \sigma(w_{fx}x_t + w_{fa}\alpha_{t-1} + w_{fc}c_{t-1} + b_f) \quad (6)$$

文本通过记忆单元的计算公式为:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot g(w_{cx}x_t + w_{ca}\alpha_{t-1} + b_c) \quad (7)$$

文本通过输出门的计算公式为:

$$o_t = \sigma(w_{ox}x_t + w_{oa}\alpha_{t-1} + w_{oc}c_{t-1} + b_o) \quad (8)$$

文本通过上述计算可获得文本编码向量的输出,具体公式为:

$$\alpha_t = o_t \cdot h(c_t) \quad (9)$$

$$y_t = \pi(w_{yx}\alpha_t + b_y) \quad (10)$$

其中, \cdot 表示点乘运算, i 表示输入门, f 表示遗忘门, f_t 表示文本在 t 时刻通过遗忘门, o 表示输出门, c 表示记忆单元, σ 表示激活函数, π 表示总的输出激活函数, g 、 h 表示记忆单元的输入和输出激活函数。

2.3 多头注意力机制

注意力机制^[21-22]是指通过计算 Q (Query) 和每个 K (Key) 之间的相似性 $\text{Sim}(Q, K)$ 以获取分配的权重,然后将分配的权重与相应的 V (Value) 值进行加权求和得到注意力权值,其中 Q 、 K 和 V 均为向量且 $K = V$ 。与单一结构的注意力机制相比,多头注意力机制是对 Q 、 K 和 V 各维度分别进行多次线性映射并对其进行拼接,以获得最终的注意力权值。

本文利用多头注意力机制^[23-24]不仅可对 WL-CNN 的低层信息和 SL-Bi-LSTM 的全局语义特征分配不同的权重,而且可进一步捕获不同位置和子空间的结构信息、深层全局语义信息以及内部结构信息。多头注意力权重的计算公式为:

$$\text{Head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (11)$$

其中, W 表示线性变化参数, Q 、 K 和 V 对应不同的 W 值。将式(11)进行多次线性映射及拼接后便可获得多头注意力的加权值,计算公式为:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{Head}_1, \text{Head}_2, \dots, \text{Head}_\pi) W^o \quad (12)$$

其中, π 表示多头注意力机制的并行层数,即该多头注意力机制包含 π 个头。

综上所述,本文使用词级卷积神经网络和句级

双向长短期记忆网络来学习词序列子空间向量和句序列深层语义信息,并通过多头注意力机制将两种

深度学习模型进行集成,实现子空间信息和句子语义信息的互补,如图 1 所示。

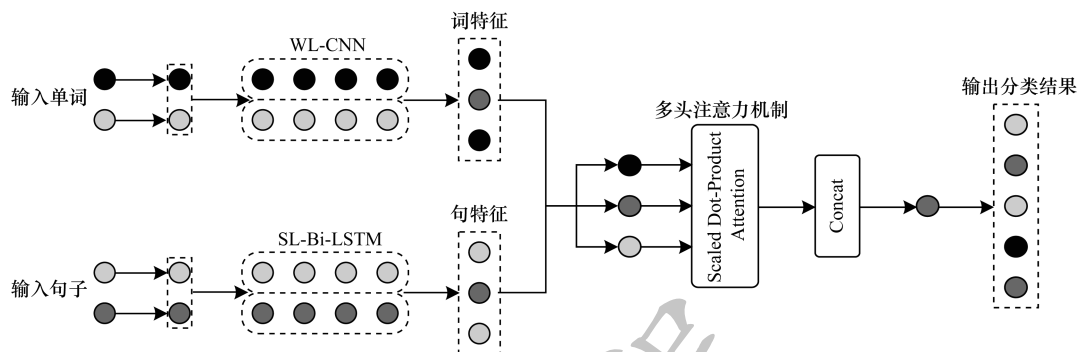


图 1 集成深度学习模型旅游问句文本分类过程

Fig.1 Classification process of tourist question texts integrated with deep learning models

可以看出,将旅游问句文本以单词形式输入词级卷积神经网络中,基于该神经网络的计算获取文本中词的低层子空间结构信息和局部深层语义信息 x^{word} ,旅游问句文本以句子的形式输入 SL-Bi-LSTM 中,通过学习该神经网络获得旅游问句文本的深层结构信息和全局语义信息 x^{sentence} 。在此基础上,利用多头注意力机制对两种深度学习模型获得的层次特征进行权值匹配,并对各注意力机制所得注意力权重进行点积拼接,通过 SoftMax 分类函数得到文本的准确分类结果 $z^1 \sim z^m$,其中 m 表示分类类别数。

3 实验结果与分析

3.1 实验数据集

本文采用 Tourism text 数据集作为实验数据集,其为自定义的基准数据集,主要来自携程、途牛、马蜂窝和同程等旅游网站,包括旅游地点、时间和人物等 6 类 10 000 条样本数据。在实验前需对该数据集进行筛选、清洗、停用词等预处理操作减少误差。为验证本文算法的有效性,从各类样本中随机抽取 60% 的样本作为训练集,剩余 40% 的样本作为测试集,并随机划分训练集中 10% 的样本作为交叉验证集(实验共进行 5 次交叉验证),数据集统计结果如表 1 所示,其中 All Doc 表示每类旅游问句文本的总数。

表 1 数据集统计结果

Table 1 Statistical results of the dataset

类别	训练集	测试集	验证集	文本平均长度	All Doc
地点	1 051	700	106	25	1 751
时间	1 200	795	120	35	1 995
实体	1 140	758	114	61	1 898
数字	1 042	694	105	74	1 736
描述	1 422	994	143	22	2 416
人物	123	81	13	2 453	204

3.2 实验方法

将本文旅游问句文本分类算法与以下主流旅游问句文本分类算法进行对比:

1) 基于 CNN^[25] 的旅游问句文本分类算法。该算法主要是对中文文本中的词采用随机初始化嵌入,并输入卷积神经网络中进行分类。为使该算法在中文问句文本分类中取得理想的分类结果,采用与文献[25]相同的随机初始化词嵌入方法,获取中文问句文本中的词向量特征。

2) 基于 Word2vec + LR^[26] 的旅游问句文本分类算法。该算法利用词向量嵌入表示每种类型的文本,并将其映射到低维空间向量中使用逻辑回归模型进行文本分类。

3) 基于 Word Embeddings + SVM 的旅游问句文本分类算法。该算法类似于基于 Word2vec + LR 的旅游问句文本分类算法,是将问句文本嵌入到低维空间向量中使用浅层机器学习模型实现文本分类。

4) 基于 LSTM 与 Bi-LSTM^[27] 的旅游问句文本分类算法。这两种算法是将训练的中文问句文本词向量按照时序依次输入网络模型,并对中文问句文本进行时序扩展和深层语义信息的捕获,实现中文问句文本的准确分类。

5) 基于自注意力网络(Self-Attention Network, SAN)的旅游问句文本分类算法。该算法结合词向量技术,将中文问句文本映射到一个低维空间向量,并通过注意力网络对关键特征分配权重实现中文文本的准确分类。

6) 基于 RNN 的旅游问句文本分类算法。该算法解决了卷积神经网络忽略全局语义信息及梯度消失等问题,进一步提高了分类算法的准确率。

7) 基于独立循环神经网络(Independently RNN, Ind-RNN)的旅游问句文本分类算法。该算法解决了 RNN 算法梯度消失及不能对长时序特征进行有效建模的问题。

8) 基于 CNN-LSTM 的旅游问句文本分类算法。该算法主要将 CNN 和 LSTM 进行串联, 在利用 CNN 捕获问句文本的空间特征后再对其进行时序建模, 以提高文本分类精度。

3.3 实验环境和参数设置

本文实验软件配置为 Python 3.6 和 Keras、Numpy 等框架, 硬件配置为 GTX1060 GPU。为确保文本分类算法的一致性, 本文对其设置初始化参数, WL-CNN 的嵌入向量设置为 512, 其具有两层卷积层且卷积核大小分别为 3×3 和 1×1 , 神经元个数为 512 和 128, 学习率为 0.000 1, 丢码率为 0.5。SL-Bi-LSTM 的嵌入向量设置为 512, 神经元个数为 256 和 64, 学习率为 0.02。在利用多头注意力机制进行集成时将 π 设置为 10 并对其分配不同的权重比例, 同时采用概率计算统计不同类别的出现概率, 实现旅游问句文本的正确分类。在前期实验中将词向量映射维度设置为 100 维, 在后期实验中随着 F-Score 的变化对词向量映射维度进行调整。

3.4 实验结果

本文采用准确率、损失率、模型训练时间和 F-Score 进行实验结果评估, 其中 F-Score 的计算公式如下:

$$F\text{-Score} = \frac{2PR}{P+R} \quad (13)$$

其中, P 表示精确率, R 表示召回率。

3.4.1 词向量映射维度对算法性能的影响

由于词向量映射维度对算法分类精度具有至关重要的作用, 因此本文通过改变词向量映射维度来验证其对旅游问句文本分类精度的影响。图 2 显示了 Tourism text 数据集在不同词向量嵌入维度下两种算法的 F-Score 对比。可以看出, 随着词向量映射维度的增加, 本文旅游问句文本分类算法的 F-Score 先快速增加, 当词向量映射维度大于 100 时 F-Score 停止增加且开始减少, 这表明过低的词向量映射维度不能较好地文本映射到低维空间, 而高维嵌入可能导致向量表示过于稀疏, 不能有效提高分类性能且会耗费更多的训练时间。

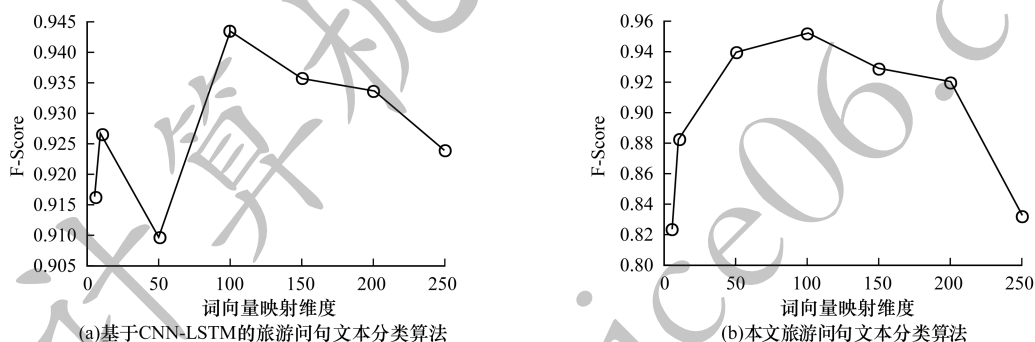


图 2 不同词向量映射维度时旅游问句文本分类算法的实验结果

Fig.2 Experimental results of classification algorithms of tourist question tests at different mapping dimensions of word vector

3.4.2 旅游问句文本分类算法的性能比较

为进一步验证本文算法的有效性, 将其与不同问句文本分类算法进行对比, 实验结果如表 2 所示, 其中加粗数据表示最优结果。

表 2 10 种旅游问句文本分类算法的性能比较结果

Table 2 Performance comparison results of ten classification algorithms of tourist question texts

分类算法	准确率	损失率	F-Score	训练时间/s
CNN ^[25]	0.949 8	0.369 7	0.925 5	15.71
Word2vec + LR ^[26]	0.933 7	0.386 7	0.900 1	15.13
Wording Embeddings + SVM ^[7]	0.941 1	0.371 9	0.913 3	19.44
LSTM	0.948 2	0.315 4	0.931 5	15.73
Bi-LSTM ^[27]	0.954 0	0.286 7	0.934 7	25.28
SAN	0.952 7	0.283 6	0.931 1	24.21
RNN	0.947 2	0.291 6	0.916 0	40.06
Ind-RNN	0.950 2	0.223 1	0.932 9	22.44
CNN-LSTM	0.961 3	0.294 8	0.950 2	17.38
本文算法	0.986 6	0.127 7	0.980 4	21.97

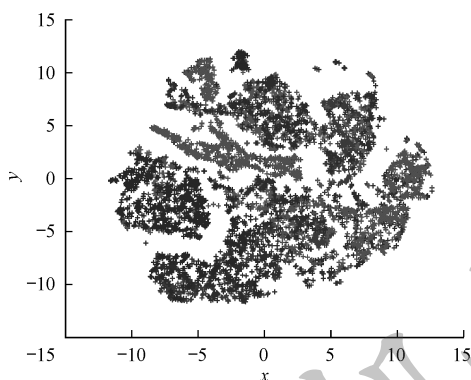
可以看出, 基于 Ind-RNN 的旅游问句文本分类算法损失率为 0.223 1, 因为 Ind-RNN 内部神经元相互独立, 且各层之间实现了跨层连接, 所以其相比基于 RNN 的旅游问句文本分类算法更好地解决了梯度消失问题。基于 Word2vec + LR 和 Wording Embeddings + SVM 的旅游问句文本分类算法虽然取得了较好的分类结果, 但其仅适用于小规模数据, 其主要原因为基于浅层机器学习模型旅游问句文本分类算法不能较好地捕获问句文本中隐藏的潜在信息, 因此其相比基于深度学习模型旅游问句文本分类算法准确率较低。基于 Bi-LSTM 的旅游问句文本分类算法的准确率和损失率均优于基于 LSTM 的旅游问句文本分类算法, 其主要原因为基于 Bi-LSTM 的旅游问句文本分类算法同时使用前向和后向 LSTM 对文本进行编码, 更好地捕获了旅游问句文本的上下文信息。与其他旅游问句文本分类算法相比, 本文算法在准确率和损失率上分别取得了 0.986 6 和 0.127 7 的最优结果, 相比基于 CNN-LSTM 的旅游问句文本分类算法提高

了0.025 3和降低了0.167 1,其主要原因为本文算法不仅能捕获旅游问句文本的深层结构信息和全局语义信息,而且可对各类特征分配不同权重,从而实现准确分类。但是其在训练时间上略显不足,其主要原因为神经网络在集成时算法的总参数量有所增加,因此导致耗时较长。

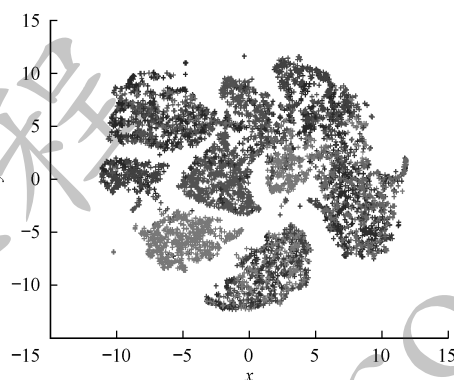
3.4.3 旅游问句文本的可视化结果

为更直观地表示旅游问句文本分类后的结果,本文使用 t-SNE 可视化工具对分类结果进行可视化

显示,得到旅游问句文本嵌入表示的可视化结果如图 3 所示。可以看出,基于 CNN + LSTM 的旅游问句文本分类算法不能较好地地区分各类旅游问句文本,其主要原因为该算法仅对旅游问句文本的局部特征和语义信息进行建模,而忽略了旅游问句文本全局信息、上下文层级关系以及文本中的词对文本的影响。与基于 CNN + LSTM 的旅游问句文本分类算法相比,本文算法能更有效地对不同旅游问句文本进行分类且分类性能更好。



(a)基于CNN-LSTM的旅游问句文本分类算法



(b)本文旅游问句文本分类算法

图 3 旅游问句文本嵌入表示的可视化结果

Fig. 3 Visualized results of embedding representation of tourist question texts

4 结束语

本文提出一种改进的旅游问句文本分类算法,通过 WL-CNN 和 SL-Bi-LSTM 捕获旅游问句文本的局部和全局特征,并有效刻画了旅游问句文本的语义信息和上下文层级关系,同时利用多头注意力机制对所捕获的特征信息分配注意力权重,实现旅游问句文本的有效分类。后续将研究旅游问句文本的细粒度分类算法,通过对旅游问句文本中含有的关键词做进一步表征,增强特征信息对旅游问题文本的表征能力并提高分类准确度。

参考文献

- [1] LIU Xinsheng, LI Kun. Design and implementation of tourism events text classification system based on BP-NN [J]. Computer and Modernization, 2011, 7(7): 192-194.
- [2] LIU Jingzhou, CHANG Weicheng, WU Yuexin, et al. Deep learning for extreme multi-label text classification [C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2017: 115-124.
- [3] XIE Yufei, LÜ Zhao. Question fine-grained classification based on semantic expansion and attention network [J]. Computer Engineering, 2019, 45(1): 171-177, 183. (in Chinese)
谢雨飞, 吕钊. 基于语义扩展与注意力网络的问题细粒度分类 [J]. 计算机工程, 2019, 45(1): 171-177, 183.

- [4] DALAL M K, ZAVERI M A. Automatic text classification of sports blog data [C]//Proceedings of 2012 Computing, Communications and Applications Conference. Washington D. C., USA: IEEE Press, 2012: 61-80.
- [5] SALLES T, ROCHA L, MOURÃO F, et al. A two-stage machine learning approach for temporally-robust text classification [J]. Information Systems, 2017, 69: 40-58.
- [6] ELARNAOTY M, FARGHALY A. Machine learning implementations in Arabic text classification [M]//SHAALAN K, HASSANIEN A E, TOLBA F. Intelligent natural language processing: trends and applications. Berlin, Germany: Springer, 2017: 295-324.
- [7] GOUDJIL M, KOUDIL M, BEDDA M, et al. A novel active learning method using SVM for text classification [J]. International Journal of Automation and Computing, 2018, 15(3): 290-298.
- [8] HUANG Xianying, XIE Jin, LONG Shuyan. Question classification method combining word vector and BTM model [J]. Computer Engineering and Design, 2019, 40(2): 384-388. (in Chinese)
黄贤英, 谢晋, 龙妹言. 融合词向量及 BTM 模型的问题分类方法 [J]. 计算机工程与设计, 2019, 40(2): 384-388.
- [9] YU Bengong, XU Qingtang, ZHANG Peihang. Question classification based on MAC-LSTM [J]. Application Research of Computers, 2020, 37(1): 40-43. (in Chinese)
余本功, 许庆堂, 张培行. 基于 MAC-LSTM 的问题分类研究 [J]. 计算机应用研究, 2020, 37(1): 40-43.

- [10] LI Chenbin, ZHAN Guohua, LI Zhihua. News text classification based on improved Bi-LSTM-CNN[C]//Proceedings of the 9th International Conference on Information Technology in Medicine and Education. Washington D. C. , USA; IEEE Press, 2018; 890-893.
- [11] ZHANG Dong, LI Shoushan, WANG Jingjing. Semi-supervised question classification with jointly learning question and answer representations [J]. Journal of Chinese Information Processing, 2017, 31(1): 1-7. (in Chinese)
张栋, 李寿山, 王晶晶. 基于问题与答案联合表示学习的半监督问题分类方法[J]. 中文信息学报, 2017, 31(1): 1-7.
- [12] LIU Guolong, XU Xiaofei, DENG Bailong, et al. A hybrid method for bilingual text sentiment classification based on deep learning [C]//Proceedings of the 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. Washington D. C. , USA; IEEE Press, 2016; 93-98.
- [13] ZHARMAGAMBEV A S, PAK A A. Sentiment analysis of a document using deep learning approach and decision trees [C]//Proceedings of the 12th International Conference on Electronics Computer and Computation. Washington D. C. , USA; IEEE Press, 2016; 93-98.
- [14] ZHANG Qing, LÜ Zhao. Domain question classification method based on topic expansion [J]. Computer Engineering, 2016, 42(9): 202-207. (in Chinese)
张青, 吕钊. 基于主题扩展的领域问题分类方法[J]. 计算机工程, 2016, 42(9): 202-207.
- [15] DUQUE A B, SANTOS L L J, MACÊDO D, et al. Squeezed very deep convolutional neural networks for text classification [C]//Proceedings of ICANN' 19. Berlin, Germany; Springer, 2019; 193-207.
- [16] WIGINGTON C, STEWART S, DAVIS B, et al. Data augmentation for recognition of handwritten words and lines using a CNN-LSTM network [C]//Proceedings of IAPR International Conference on Document Analysis & Recognition. Washington D. C. , USA; IEEE Press, 2018; 1-10.
- [17] ADAK C, CHAUDHURI B B, BLUMENSTEIN M. Offline cursive Bengali word recognition using CNNs with a recurrent model [C]//Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition. Washington D. C. , USA; IEEE Press, 2016; 429-434.
- [18] NIU Jianwei, SUN Mingsheng, MO Shasha. Sentiment analysis of Chinese words using word embedding and sentiment morpheme matching [C]//Proceedings of International Conference on Collaborative Computing: Networking, Applications and Worksharing. Berlin, Germany; Springer, 2018; 3-12.
- [19] ZHANG Yangsen, ZHENG Jia, JIANG Yuru, et al. A text sentiment classification modeling method based on coordinated CNN-LSTM-attention model [J]. Chinese Journal of Electronics, 2019, 28(1): 124-130.
- [20] LI Lishuang, JIANG Yuxin. Biomedical named entity recognition based on the two channels and sentence-level reading control conditioned LSTM-CRF [C]//Proceedings of 2017 IEEE International Conference on Bioinformatics and Biomedicine. Washington D. C. , USA; IEEE Press, 2017; 380-385.
- [21] ZHOU Botong, SUN Chengjie, LIN Lei, et al. LSTM based question answering for large scale knowledge base [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2018, 54(2): 286-292. (in Chinese)
周博通, 孙承杰, 林磊, 等. 基于 LSTM 的大规模知识库自动问答[J]. 北京大学学报(自然科学版), 2018, 54(2): 286-292.
- [22] WU Lin, WANG Yang, LI Xue, et al. Deep attention-based spatially recursive networks for fine-grained visual recognition [J]. IEEE Transactions on Cybernetics, 2019, 49(5): 1791-1802.
- [23] ALKHOULI T, BRETSCHNER G, NEY H. On the alignment problem in multi-head attention-based neural machine translation [EB/OL]. [2019-08-03]. <https://arxiv.org/abs/1809.03985?context=cs>.
- [24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA; ACM Press, 2017; 6000-6010.
- [25] GU Chengwei, WU Ming, ZHANG Chuang. Chinese sentence classification based on convolutional neural network [J]. IOP Conference Series: Materials Science and Engineering, 2017, 261: 12-28.
- [26] PRANCKEVICIUS T, MARCINKVICIUS V. Application of logistic regression with part-of-the-speech tagging for multi-class text classification [C]//Proceedings of the 4th Workshop on Advances in Information, Electronic and Electrical Engineering. Washington D. C. , USA; IEEE Press, 2016; 9-18.
- [27] SHIH C H, YAN B C, LIU S H, et al. Investigating Siamese LSTM networks for text categorization [C]//Proceedings of 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Washington D. C. , USA; IEEE Press, 2017; 641-646.

编辑 陆燕菲