文章编号:1007-757X(2024)02-0018-03

基于词向量融合的建筑文本分类方法研究

胡少云， 翁清雄
（中国科学技术大学，管理学院，安徽，合肥 230026）

摘 要：由于建筑领域问题包含复杂多样的领域专有术语，常见的文本分类算法在建筑领域问题分类上难度较大。为提高建筑领域问题的分类性能，提出一种基于融合 RoBERTa 和 Word2Vec 的建筑文本分类算法。实验结果表明：在建筑领域问题数据集上，准确率达到 91.59%，分类性能较好；在通用数据集上，准确率均高于 SVM、CNN 等模型。

关键词：文本分类；预训练语言模型；句向量；深度学习；问答系统
中图分类号：TP391.1 文献标志码：A

Research on Architectural Text Classification Method Based on Word Vector Fusion

HU Shaoyun, WENG Qingxiong
(School of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract: Due to the complexity and variety of domain-specific terms in architectural questions, the common text classification algorithms are more difficult to classify architectural questions. In order to improve the classification performance of questions in the architectural field, this paper proposes an architectural text classification algorithm based on the fusion of RoBERTa and Word2Vec. Experimental results show that the accuracy rate of the proposed method reaches 91.59% on the construction domain problem dataset, and the classification performance is better, and on general data sets, the accuracy rate is higher than that of SVM, CNN and other models.

Key words: text classification; pretrained language model; sentence vector; deep learning; question-answering system

0 引言

人工智能等技术对知识培训产生深远影响，智能问答系统因其实时交互、高效便捷、解答精准等优势，在建筑知识培训中得到广泛应用。文本分类作为智能问答系统的重要支撑技术之一，成为研究的热门领域。目前，实现文本分类常见的两类方法分别是基于规则词典方法和基于机器学习方法[1]。随着文本数据海量增加，以往常用的逻辑回归、SVM 等传统机器学习方法的性能已经无法满足用户需求。深度学习模型因其在海量数据处理上效果显著，成为文本分类任务的主流方法[2]。

在基于深度学习模型文本分类研究方面，国内外研究团队取得了很好的研究成果。JOHNSON 等[3]提出了一种深度卷积神经网络结构，可以有效提取文本长距离关联特征，提高文本分类率。PENG 等[4]考虑到句子间、定义字词之间的句法关系和词义关系呈现图结构，提出了一种基于 Graph-CNN 的深度学习模型。CUI 等[5]创建了一系列包括 RoBERTa-wwm-ext 在内的中文预训练语言模型，并在 10 个中文 NLP 任务上评估模型性能，该系列模型在文本分类等基线任务上效果明显。尽管深度学习在通用语料的文本分类任务上取得了优异效果，但由于特定领域语料在词汇和句子结构上与通用语料有很大不同，提取文本特征并不完备精确，因此在该类语料上分类效果较差。建筑领域语料作为具有代表性的特定领域语料，包含大量分布不规律且难以收集的专业术语，极大地提高了文本分类的难度。针对该问题，本文提出一种基于词向量融合的建筑文本分类方法。

1 研究方法

1.1 研究方法简述

该算法分为数据预处理、句向量生成、深度学习模型搭建等 3 个阶段。首先，数据预处理。先构建建筑术语词典，基于词典对语料进行预处理，将处理后的词表分别送入 Word2Vec 模型和 RoBERTa 模型。其次，句向量生成。将 Word2Vec 模型生成的词向量加权均值，生成基于 Word2Vec 的句向量；同时引入特定目标函数，对 RoBERTa 模型进行建筑掩码语言建模(AriMLM)任务训练，接着利用 Sentence-Bert 框架提取生成基于 RoBERTa 模型的句向量。最后，深度学习模型搭建。将上述两类句向量进行拼接，将

基金项目：国家自然科学基金重点国际(地区)合作交流项目(7191001010)
作者简介：胡少云(1973－)，男，博士研究生，高级工程师，研究方向为智能推荐、人工智能、人才测评与人才发展；
翁清雄(1981－)，博士，教授，研究方向为人才测评与人才发展、创业行为与创业发展、职业心理与行为、情绪管理、领导行为。

拼接向量作为输入,送入由 TextCNN、BiLSTM 等网络层组建的深度学习模型进行训练。本文算法框架如图 1 所示。

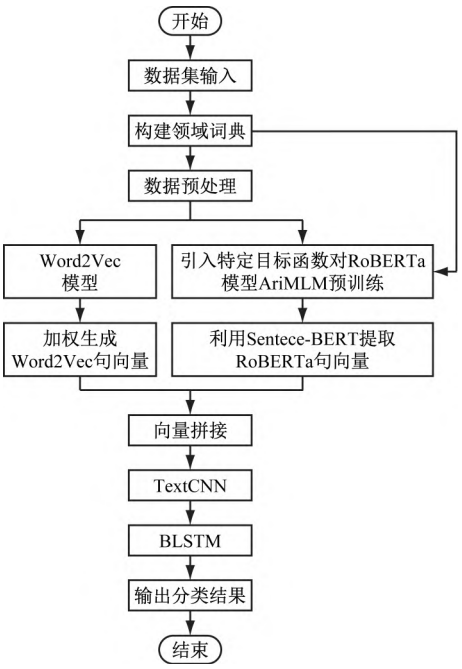

图 1　本文算法框架图

1.2　数据预处理

建筑领域问题语料中包含的术语分布规律性不强,因此建筑语料分词质量较差[6]。为了解决该问题,首先构建建筑术语词典 V,采用 LTP 工具[7]加载词典 V,对建筑问题语料进行分词,得到问题分词表 S。

1.3　句向量生成

1.3.1　Word2Vec 句向量生成

将问题分词表 S 送入 Word2Vec 模型中进行训练,获取基于 Word2Vec 模型的词向量。采用加权的方法,生成基于 Word2Vec 的句向量 WSec。

1.3.2　RoBERTa 句向量生成

基于 RoBERTa 句向量的生成过程分为两部分,采用哈工大提出的 RoBERTa-wwm-ext[5]模型,引入特定预训练目标函数,结合建筑术语词典 V,对模型进行 AriMLM 任务训练。利用 sentence-bert 框架对训练后的模型生成建筑语料句向量。RoBERTa 模型由 Embeding 层和多层 Transformer 层组成,与 BERT-wwm 结构基本一致。模型架构如图 2 所示。将单句 L 送入 Embedding 层,提取词嵌入向量(Token Embedding)、分割嵌入向量(Segment Embedding)、位置嵌入向量(Position Embedding)分别表示词在字词表的位置、词所属句子的信息、词的位置信息。上述 3 类向量求和得到输入序列 X。

Transformer 层的结构如图 3 所示。该层是 RoBERTa 模型的重要部分,该层由多个 Encoder 层组成,Encoder 层由多头注意力机制(multi-head attention)、残差归一化网络(add&norm)、前馈神经网络(feed forward neural network)

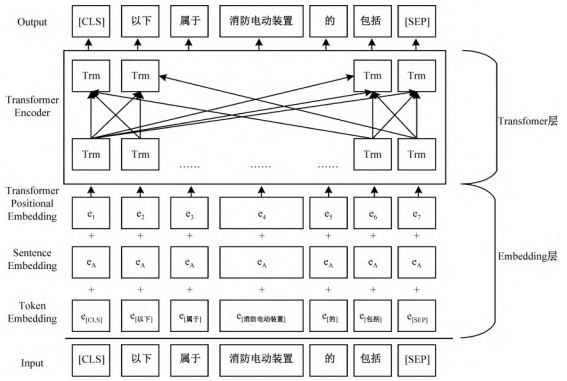

图 2　RoBERTa-wwm-ext 模型架构图

3 种网络结构层组成。multi-head attention 层作为 transformer 的核心,利用注意力机制计算词和词之间的关联度矩阵,调节词权重。

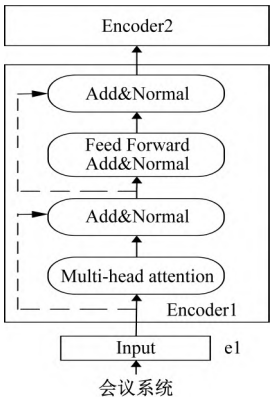

图 3　Transformer 层的架构

为获取适合建筑文本分类的语言表示特征,提出一种 AriMLM 方法。该方法掩码方式与对所有词以 15% 的概率进行掩码语言建模方法的不同之处在于将建筑词典 V 的信息引入 BERT 预训练目标中,在建模时,赋予词典 V 中的词更高掩码频率。如在表 1 中,输入语句“以下关于火灾探测器描述不正确的是”,提取在词典 V 中的词“火灾探测器”作为词典 V,那么设置该词的掩码概率为 45%,其他词的掩码概率则下降为 10.7%,总体掩码概率仍然保持在 15%,训练的其他部分与原来的 RoBERTa 模型的训练过程一致。将问题分词表 S 送入模型经过 AriMLM 训练后,得到面向建筑领域的预训练语言模型。

BERT 网络结构的缺点之一是没有单独计算独立的句嵌入向量,导致从 BERT 网络中提取句嵌入向量较难、句向量性能较差。为解决上述问题,采用 sentence-bert 框架。该框架基于孪生和三元网络结构,将预训练语言模型的输出进行池化操作,得到固定大小的句子嵌入。

1.4　深度学习模型搭建

为了提取不同个数相邻词的关联特征,模型定义高度 h 分别为 3、4、5 的不同类型卷积核,对句向量 Sv 进行卷积操作。由于卷积核提取的数据特征过多,会引起模型过拟合,因此对所有卷积核输出结果进行最大池化操作,并将结果连接为向量,获得的特征集合为 $C^{(2)}$。为了提取文本中的双向时序关系,在原有的卷积层网络后连接一层相反方向的双向

LSTM 网络。同时，为增强模型的泛化能力，模型添加一层嵌入 Dropout 方法的全连接层，用于对数据进行加权运算后得到最终特征向量 Vec，将 Vec 送入 Softmax 函数获取建筑问题分类的预测标签 Y。

2 实验结果与分析

2.1 实验设置

实验平台采用 Intel i7 3.80 GHz CPU、GeForce 3090 GPU、内存 16 G 的服务器，采用简单易用的 PyTorch 深度学习框架。实验评估指标采用准确率 Accuracy。

实验的数据集分为两部分，验证算法有效性的数据集是建筑问题集。该数据集根据《弱电工职业技能培训教材》采集整理出 30 082 条建筑施工人员频繁查询的问题，问题类型有 4 类，如表 1 所示。验证算法通用性的数据集有 3 类，如表 2 所示。数据集 A 取自今日头条新闻语料，抽取了军事、房产、体育等 3 类新闻若干条作为数据集，数据集 B 是外卖评价数据集，数据集 C 是酒店评价数据集。

表 1 建筑语料数据集

序号	问题	问题类型
1	火灾自动报警系统包含以下哪几种设备	定义
2	以下光警报器应安装距地距离不正确的包括	安装
3	以下不是报警控制器提示故障的原因包括	调试
4	停车库(场)管理系统规定要求表述不正确的包括	功能

表 2 通用语料数据集

数据集	样本数量	标签类型
A	80 224	3
B	11 987	2
C	7766	2

2.2 实验结果及分析

验证算法有效性实验中，通过固定其他超参数，更改特定参数的数值，设置最佳参数的情况进行实验，验证模型的训练次数 Epoch、学习速率、Dropout 值等超参数对模型分类性能的影响。

训练次数 Epoch 对模型性能影响极大：训练次数过少会导致模型过拟合；训练次数过多则会降低模型的泛化能力。图 4 显示训练次数在 1～15 时准确率的变化趋势。由图 4 可知，在训练 12 次后，准确率 Accuracy 保持稳定，并在 15 次时达到最高点，因此设置实验 Epoch 最佳值为 15。改变学习速率的大小在一定程度上影响模型性能。当速率过大时，模型难以收敛；而当速率过小时，模型无法学习或收敛速度过慢。图 5 反映了模型在训练集和测试集上受学习速率的影响情况：当学习速率是 0.6 时，模型曲线到达最高点，训练准确率和测试准确率均达到 91.59%以上，模型分类性能最佳，之后模型准确率随着学习速率变大而下降。为此，设置模型学习速率为 0.6，保持模型处于最佳性能。Dropout 值表示训练模型时模型神经元的失活比率。Dropout 技术将有利于目标函数寻找到优化解，在一定程度上增强模型的泛化性，加快模型训练速度。图 6 显示 Dropout 值对模型在训练集上的变化趋势：当 Dropout 小于 0.4 时，模型准确率会提升，并且模型准确率会波动；当 Dropout 大于 0.4 时，模型准确率则会下降。因此，设置模型的 Dropout 值为 0.4。

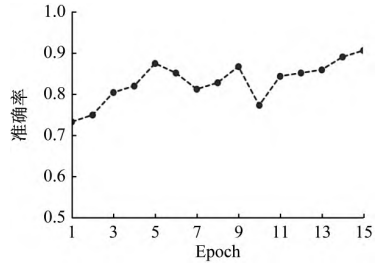

图 4 Epoch 的影响图

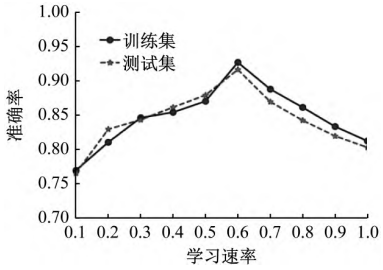

图 5 学习速率的影响图

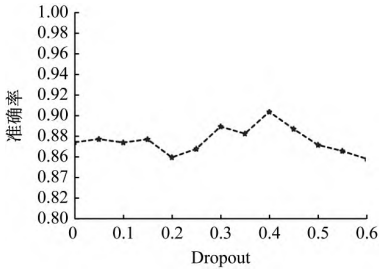

图 6 Dropout 的影响

表 3 不同模型实验结果

模型	A		B		C	
	Train Acc	Test Acc	Train Acc	Test Acc	Train Acc	Test Acc
SVM	0.9645	0.9613	0.8632	0.8603	0.8849	0.8571
GaussianNB	0.9401	0.9400	0.8257	0.8182	0.8086	0.7921
TextRNN_Att	0.9706	0.9489	0.8338	0.8084	0.9000	0.8874
TextRCNN	0.9436	0.9235	0.8426	0.8124	0.8991	0.8900
本文模型	0.9757	0.9773	0.8764	0.8723	0.9375	0.9311

(下转第 25 页)

从表 2 可以看出,水平扫描方向的误差在 5%以内。随着信号频率的增大,信号的幅值在减小,说明系统受模拟带宽的限制,当频率高到一定程度时信号出现了衰减。

5　总结

本文研究了掌上多功能数字示波器系统,利用 STM32 芯片内部集成定时器、ADC、DAC、DMA 等模块,结合外部信号处理电路,实现了数字示波器、信号发生器和电子相册的功能。通过滤波和插值算法将各项指标误差率控制在 5%之内。实验测试结果表明,本文系统能够输出简易类型信号,可稳定测量信号波形且有电子相册功能。本文多功能示波器的成功研制为新型示波器的研究提供了新的设计方法和思路。

参考文献

[1] 田祥祥. 多功能虚拟示波器模块的硬件设计与实现[D]. 成都:电子科技大学,2019.

[2] 刘素贞,魏建,李华,等. 基于 STM32 的多功能虚拟示波仪的设计[J]. 实验技术与管理,2018,35(1):152-156.

[3] 季小榜,孙雷明. 基于 STM32F103RCT6 的虚拟示波器硬件电路设计[J]. 河南工程学院学报(自然科学版),2020,32(1):37-40.

[4] 刘东,曾仕鹏,刘雪敬. 基于 STM32 的便携式示波器的设计[J]. 科技创新与应用,2017(12):77.

[5] 李璐,李腾飞,李飞飞. 基于单片机和 FPGA 的数字示波器的设计[J]. 电子设计工程,2011,19(18):78-81.

[6] 狄飞. 基于 FPGA 与 STM32 的手持式数字示波器设计[D]. 绵阳:西南科技大学,2018.

[7] 张华忠. 基于 STM32 的便携式数字示波器设计与实现[J]. 现代计算机(专业版),2017(19):45-47.

[8] 徐健,唐胤. 基于 STM32 的便携式数字示波器设计[J]. 电子设计工程,2019,27(14):139-143.

[9] 孙丽娟. 便携式数字示波器关键技术研究[D]. 西安:西安电子科技大学,2018.

[10] 邹小航,张心怡,李刚. 基于 STM32 单片机的函数信号发生器[J]. 电子制作,2019(9):15-17.

[11] 包欢欢. 基于虚拟仪器技术的虚拟示波器设计[J]. 信息与电脑(理论版),2019(2):64-66.

[12] 郭宏伍,马东吉,黄金成,等. 基于 ARM 的便携式数字示波器设计[J]. 科技视界,2018(12):243-244.

(收稿日期:2022-07-12)

(上接第 20 页)

为进一步验证算法的通用性,将本文模型与 SVM、GaussianNB、TextRNN_Att、TextRCNN 等模型在通用数据集 A、B、C 上进行对比测试。由表 3 可以看出:本文提出模型在数据集 A 上的准确率为 97.57%,在数据集 B 上的准确率为 87.64%,在数据集 C 上的准确率为 93.75%,均取得最高准确率,分类性能最佳;在数据集 B 和 C 上,分类性能均远超其他模型;在数据集 A 上,模型在测试集上的准确率均高于其他模型,说明该模型具有较强的泛化能力。经分析,模型经过特定预训练,获得了接近建筑领域问题的语言表示特征,通过 sentence-bert 框架生成更接近建筑术语真实分布空间句向量,与 Word2Vec 加权得到的句向量拼接后,句嵌入向量效果大大提升。该模型经过 CNN 网络层和 Bi-LSTM 网络层抽取了文本的局部特征和时序特征,极大地提升了模型学习综合特征的学习能力,之后采用 Dropout 技术,大大提高了模型的泛化能力。

3　总结

本文提出的 AS-WV-DL 方法在建筑问题分类上准确率较高,在通用数据集上的分类性能超越其他模型。但由于算法中涉及语言模型预训练和不同的深度神经网络融合,训练参数过多,容易发生过拟合风险和训练时间过长等问题。因此下一步的研究工作将深入探究文本语义分析和模型压缩方向。

参考文献

[1] LI Q,PENG H,LI J X,et al. A Survey on Text Classification: From Shallow to Deep Learning[EB/OL]. 2020: arXiv: 2008.00364. http://arxiv.org/abs/2008.00364.Pdf

[2] MINAEE S,KALCHBRENNER N,CAMBRIA E,et al. Deep Learning: Based Text Classification: A Comprehensive Review[J]. ACM Computing Surveys,2021,54(3):62.

[3] 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). Vancouver,Canada. Stroudsburg,PA,USA: Association for Computational Linguistics,2017:562-570.

[4] PENG H,LI J X,HE Y,et al. Large-scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN[C]// Proceedings of the 2018 World Wide Web Conference. ACM,2018: 1063-1072.

[5] CUI Y M,CHE W X,LIU T,et al. Pre-training with Whole Word Masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2021(29):3504-3514.

[6] 张鑫. 面向建筑领域的中文分词方法研究[D]. 北京:北京建筑大学,2022.

[7] CHE W X,FENG Y L,QIN L B,et al. N-LTP: An Open-source Neural Chinese Language Technology Platform for Chinese[EB/OL]. 2020: arXiv: 2009.11616. http://arxiv.org/abs/2009.11616.Pdf.

(收稿日期:2023-08-23)