基于提示学习混合模型的学术论文自动分类研究*

刘爱琴　贺玉斌　马茹茹
（山西大学经济与管理学院，太原 030006）

摘要：学术论文分类在知识管理、学术交流、研究导向和学术评估等方面都具有重要的意义。基于深度学习模型构建学术论文自动分类系统，相较于现有的文本分类方法，该系统融合提示学习思想，可较好地缩小预训练模型与下游任务的差距。结果表明，该系统较好地提高了文本分类性能和规范性，为科研工作者提供了更好的管理、利用和挖掘信息的方式。

关键词：学术论文；提示学习；自动分类
中图分类号：G353.1　DOI：10.3772/j.issn.1673-2286.2024.04.009
引文格式：刘爱琴，贺玉斌，马茹茹. 基于提示学习混合模型的学术论文自动分类研究[J]. 数字图书馆论坛，2024，20（4）：74-80.

随着学术论文数量的不断增多，科研工作者在日常科研活动中需要花费更多的时间来检索所需论文。同时，学术论文数量的激增对论文分类的效率和精确度提出了新的挑战[1]。构建论文分类系统有助于快速准确地组织和检索学术文献，促进科研工作的高效开展，为科研工作者提供更高效、便捷的学术资源。深度学习模型在文本分类领域中的表现愈加出色，但采用模型进行文本分类时，由于训练内容与实际下游任务之间存在差距，需要对模型进行微调[2]，微调过程中必然会损失一些预训练过程中的知识，从而影响模型效果[3]。而提示学习不使用“预训练—微调”的模式，就可以使预训练模型满足下游任务需求，缩小两者之间的差距，提升预训练模型使用效率[4]。因此，本文提出了基于提示学习的混合模型，通过加入模板和构建标签词表的方式，将下游任务建模为语言模型的掩码生成问题[5]。同时论文的标题、关键词和摘要通常被认为是论文的核心元素，它们共同提供了关于论文内容和主题的多源信息。对多源信息特征进行拼接组合，选取最佳特征组合，可以进一步增强模型的分类效果[6]。结合深度学习模型，融合提示学习思想与论文多源信息，对学术论文进行自动分类，可以实现更加准确的学术论文分类，满足相关研究方向学术工作者的需求。

1　相关研究

学术论文分类属于自然语言处理的文本分类任务，而在文本分类中，特征提取是非常关键的一步，直接影响分类模型的性能。相关研究涉及文本分类方法、论文分类方法和文本特征提取方法。

1.1　基于深度学习模型的文本分类方法

目前，基于深度学习模型的一系列方法已经能较好

收稿日期：2024-01-02
*本研究得到国家社会科学基金项目“中文学术领域命名实体的知识图谱构建研究”（编号：18BTQ072）资助。

地应用于文本分类领域。齐斌等[7]对稀疏表示的分类算法的限制性和复杂性进行了分析，为了避免表示系数扰乱结果，引入了约束条件，并对分类器的设计提出了一些可行建议。程彭圣男等[8]提出以BERT-BiLSTM模型为基础的学习模型，采用混合深度学习模型实现巡检文本智能分类，且证明该模型与单一主流深度学习模型TextCNN、BERT、BiLSTM相比性能更优。穆建媛等[9]提出了一种基于提示学习的短文本分类方法，在训练数据较少的情况下表现较好。

1.2　论文分类方法研究

在对论文分类方法的研究中，不同学者对于论文分类的角度有着不同的见解与看法。章成志等[10]构建了6种不同的分类模型，对学术论文使用的研究方法进行自动分类，并对分类结果进行评估。杨秀璋等[11]提出了一种基于多视图融合的论文自动分类方法，考虑论文标题、关键词、摘要3个要素的互补性和协调性，实现对海量论文的自动分类。王末等[12]用深度学习语言表征模型学习论文句子表达，对学术论文语步结构分类方法进行了研究。李湘东等[13]提出期刊论文语言规范，该规范专业性强，平均文本长度为18个字符，符合短文本的条件，可作为短文本分类的语料库，使分类结果更加系统、严谨。张智雄等[14]基于层次分类思想设计了多层次的分类器集群的科技文献自动分类引擎系统，并实现了科技文献自动分类中的高质量数据获取、多层次分类器构建、处理流程优化和开放接口调用。

1.3　文本特征提取方法

一些学者通过改进文本特征提取的方式，提升算法或模型分类准确率。Thirumoorthy等[15]提出了一种以词频分布测度为基础的特征选择方法并对比了其与其他分类方法，证明该特征选择方法具有更好的效果。Sadr等[16]利用BERT模型对文本数据开展向量编码操作，并采用卷积网络提取局部特征以提升模型分类效果，该方法具有明显效果。李杰等[17]使用预训练模型来提取文本特征表示，以提高模型任务处理的性能。

基于预训练模型的文本分类方法已取得一系列进展，但仍存在预训练模型与下游任务关联性不强的问题。本文将中国知网中的期刊数据作为训练集，嵌入提示学习方法，设计了一种改进词向量提取和循环神经网络的深度学习论文自动分类系统，较好地缩小了预训练模型与下游任务之间的差距。该系统首先将BERT与BiLSTM作为混合分类模型，在BERT模型输出后链接BiLSTM模型，提取论文标题、摘要与关键词特征并接入全链接层，随后利用Softmax函数进行分类。实验证明，该方法可以有效提高学术论文自动分类系统的准确率。

2　理论基础

2.1　BERT语言模型

BERT语言模型是一种基于Transformer的预训练模型，研究中BERT模型输出层利用多头注意力机制对信息进行整合，详见式（1）。

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{1}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别为3种不同的向量表示；d为缩放因子，用于控制映射范围。

注意力机制层将由BERT模型输入的序列分别乘以对应权重，并转换为不同的向量表示，以保证生成的词向量融合更多的上下文语义。

2.2　LSTM

在文本分类领域中，典型的循环神经网络模型存在梯度问题，无法准确表示文本信息。而LSTM作为循环神经网络的变形结构，具备长期记忆能力，较好地解决了循环神经网络梯度爆炸的问题。LSTM中每个时刻不同单位的存在形式分别见式（2）~（4）。研究采用两层LSTM作为BiLSTM层的前向传播通道和后向传播通道，分别提取上文语义特征和下文语义特征。同时将两个通道所提取的特征融合，得到上下文语义特征。

$$f_t=\sigma\left(W_f\ h_{t-1},x_t+b_f\right) \tag{2}$$

$$i_t=\sigma(W_i\ h_{t-1},x_t+b_i) \tag{3}$$

$$C_t=f_tC_{t-1}+i_t\tilde{C}_t \tag{4}$$

式中：t为时刻；f为忘记门信号，σ为Sigmoid函

数，h为输出部分，x为序列输出，W、b为模型参数；i为输出门信号；C为Cell输出信号，$\tilde{C}$为输入信号，$\tilde{C}_t=\tanh\left(W_C\left[h_{t-1},x_t\right]+b_C\right)$。$h_t=o_t\tanh\left(C_t\right)$，$o$为模型参数。

2.3 提示学习

提示学习是一种学习方法，其在不显著修改预训练语言模型结构和参数的情况下，通过向输入层添加提示信息，将下游任务转变为文本生成任务，具体描述见表1。

表1 提示学习基本框架

输入	输出
[CLS]待分类文本[SEP]本文的主题是[MASK][MASK][SEP]	[CLS]教育[SEP] [CLS]经济[SEP]

本文选择完形填空提示引入提示信息，以保证中文短文本分类任务信息输入形式表达流畅。在所构建的系统中，具体操作是在文本前插入一个[CLS]符号，并将与该符号相对应的输出向量用作整个短文本的语义表示，用于分类。在原始文本中插入模板之后，将其作为新的输入数据输入预训练的BERT模型。

2.4 指标评价

因本研究涉及多分类任务，需要评估整体的分类性能，除考察模型基本的准确率（Accuracy）、精确率（Precision）以及召回率（Recall），引入宏（Macro）平均评价指标与微（Micro）平均评价指标。宏平均指标Macro-F1按类别计算，取平均的F1值作为最后值，在样本不均衡且各个类别同等重要的情况下可以使用；微平均指标Micro-F1注重样本真实分布，在局部评估单个二分类模型性能的基础上，将F1值合并，以此评价模型效果。指标计算公式如式（5）~（10）所示。

$$P_{\text{Macro}}=\frac{\sum_{j=1}^{n}P_j}{n} \tag{5}$$

$$R_{\text{Macro}}=\frac{\sum_{j=1}^{n}R_j}{n} \tag{6}$$

$$S_{\text{Macro\_F1}}=\frac{2\times P_{\text{Macro}}\times R_{\text{Macro}}}{P_{\text{Macro}}+R_{\text{Macro}}} \tag{7}$$

$$P_{\text{Micro}}=\frac{\sum_{j=1}^{n}a_j}{\sum_{j=1}^{n}a_j+\sum_{j=1}^{n}b_j} \tag{8}$$

$$R_{\text{Micro}}=\frac{\sum_{j=1}^{n}a_j}{\sum_{j=1}^{n}a_j+\sum_{j=1}^{n}c_j} \tag{9}$$

$$S_{\text{Micro\_F1}}=\frac{2\times P_{\text{Micro}}\times R_{\text{Micro}}}{P_{\text{Micro}}+R_{\text{Micro}}} \tag{10}$$

式中：P为精确率；R为召回率；$S_{\text{Macro\_F1}}$、$S_{\text{Micro_F1}}$分别为Macro-F1、Micro-F1指标；a为被正确识别的正样本数量；b为误报的负样本数量；c为漏报的正样本数量。

3 模型构建

构建的基于BERT-BiLSTM的提示学习混合模型基本结构（见图1）由预处理、特征选择、特征关联、文本表示、分类器构造、文本分类、结果评价等模块构成。图1中，e_1、e_2为离散提示，$\boldsymbol{T}_n$为经BERT模型输出的Embedding序列，$\vec{h}_n$、$\overleftarrow{h}_n$、$\boldsymbol{h}_n$分别为前向LSTM、后向LSTM、合并层的输出。

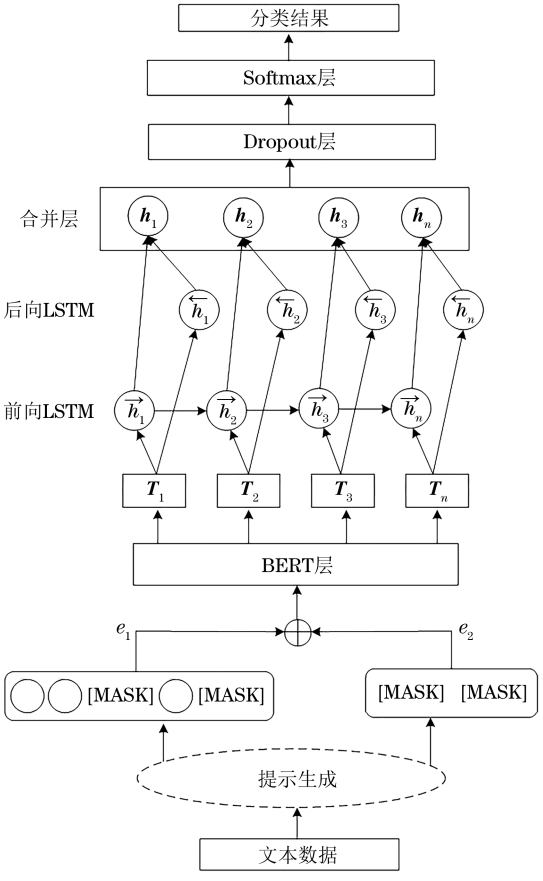


图1 基于提示学习的BERT-BiLSTM混合模型基本结构

对于给定的经过预处理的文本序列输出，该模型利用融入多头注意力机制的BERT模型对文本数据进行特征处理、编码，完成特征构造。编码层采用多头注意力机制和残差链接机制结构的多层Transformer单元堆叠，进行归一化。模型第一层的输出是融合相关文本中各个位置的字向量、文本向量和位置向量的语义信息的向量表示，生成的向量尽可能结合了上下文语义信息，更有利于提升模型最终准确率。将这些数据用于训练构建的BiLSTM模型，表示输入层的数据，用于学习文本序列的上下文信息和全局特征，文本表示部分利用BiLSTM进行具体文本建模。将基于注意力机制的标签特征应用于BiLSTM层输出的特征向量，增强模型分类的精准度。

论文分类系统（见图2）后台主要包括6个过程模块：语料检测模块、预处理模块、分词序列模块、特征构造模块、文本建模模块、类别标签模块。

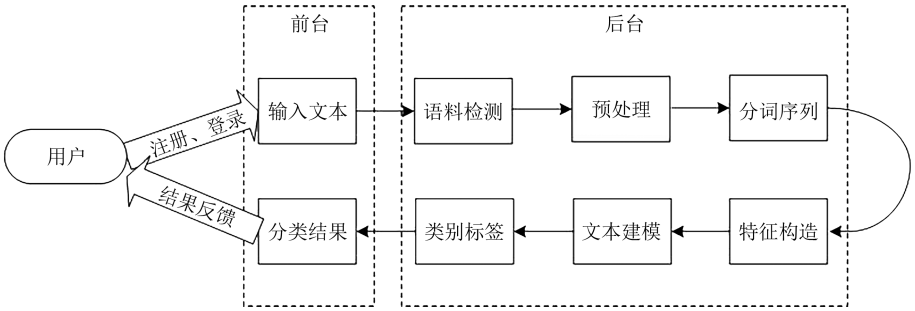


图2　论文分类系统模块图

4　实验结果与分析

将中国知网作为语料库，利用网络爬虫软件从其数据库中获取数据集，依照《中国图书馆分类法》划分的文献类别选取实验语料集，所选取的集合分别由经济、计算机、环境、教育4类论文构成，语料集包括论文的标题、摘要、关键词信息。共采集2019—2023年部分相关数据32 058条，按照8∶1∶1的比例分为训练集25 672条、验证集3 202条、测试集3 184条，实验测试样本如表2所示。混合深度学习模型参数设置如表3所示。

表2　实验数据集

单位：条

文本类别	数据量
经济	902
计算机	804
环境	799
教育	679

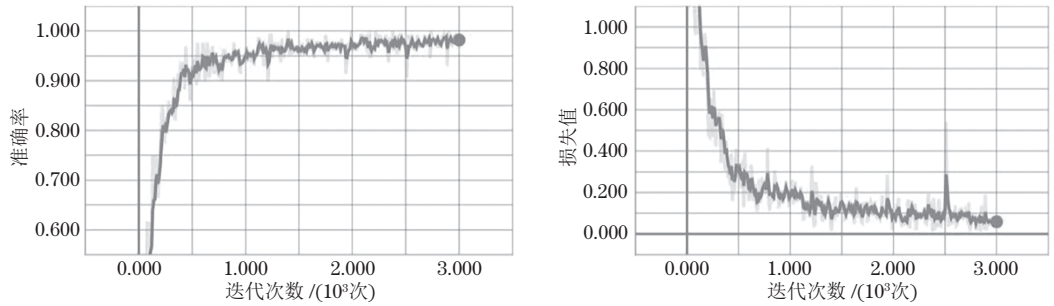
在使用训练集预训练模型的过程中，发现随着迭代次数的增加，在损失不断变小的情况下，相应的准确率不断提升，最终向1靠近。在训练过程中，损失不断减小且准确率不断提高是模型收敛的表现，收敛意味着模型学习了数据的模式和特征，并且在验证集中表现良好。准确率、损失曲线如图3所示。

表3　模型参数设置

参数	设置
Embedding	512
Transformer layer	12
BiLSTM hidden size	512
Epochs	5
Batch size	128
Pad size	32
Dropout	0.2
Optimizer	Adam
Learning rate	1.0×10^{-4}

4.1　提示学习模板对于分类结果的影响

训练完成后，将测试集中的数据输入所构建的混合模型进行自动分类。加入提示学习模板前后，测试集中的3 184条数据在BERT-BiLSTM模型上的分类结果混淆矩阵（Confusion Matrix）如图4~5所示，其中：色块颜色深浅表示各个类别预测的数据量的大小，数值表示预测的数据量，纵坐标代表论文实际类别，横坐标代表模型预测类别。对比实验结果如表4所示，可见加入提示学习模板后的模型的准确率、Macro-F1值、Micro-F1值均有提升，分类结果更加准确。



(a)准确率　　(b)损失

图3　训练准确率、损失曲线

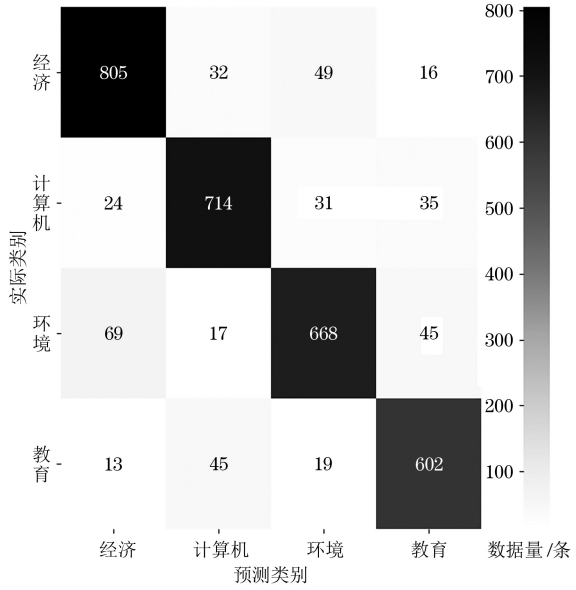


图4　未加入提示学习模板的测试集结果

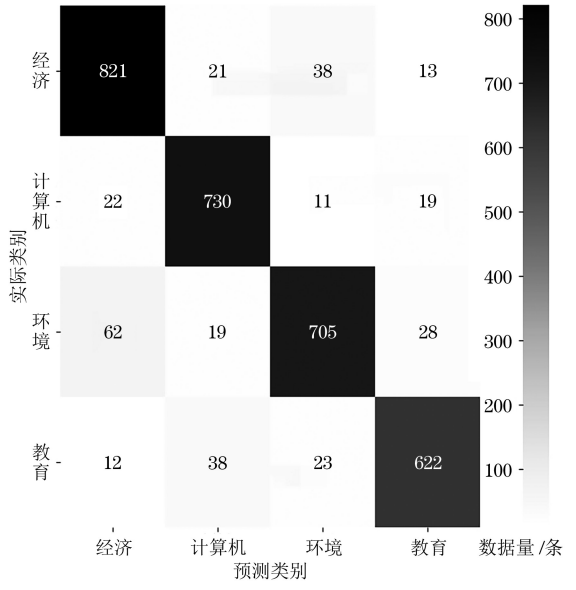


图5　加入提示学习模板的测试集结果

表4　加入提示学习模板前后的对比实验结果

类型	准确率	Macro-F1值	Micro-F1值
加入提示学习模板前	0.875 9	0.875 4	0.875 9
加入提示学习模板后	0.903 9	0.903 8	0.903 9

4.2　不同模型对于分类结果的影响

为验证所提出模型的有效性，设置了多组对照实验。首先通过类比实验验证BERT-BiLSTM模型与Word2vec-LSTM、BERT、ERNIE、BERT-LSTM 4种传统模型分类方法的效果，不同模型在经济类数据集上的预测准确率如表5所示。

表5　不同模型预测结果准确率

模型名称	准确率
Word2vec-LSTM	0.805 9
BERT	0.835 7
ERNIE	0.829 1
BERT-LSTM	0.857 7
BERT-BiLSTM	0.903 9

实验结果显示：Word2vec-LSTM模型准确率较低，这可能是因为其对文本特征的提取能力相对较弱，从而影响分类效果。BERT和ERNIE模型基于预训练的Transformer结构，在对文本进行编码时具有更好的表现，但单独使用未能达到最佳性能。ERNIE模型的准确率略低于BERT模型，这可能是因为ERNIE的优势在于对实体关系建模，而论文分类任务更侧重于全局语境的捕捉而非实体关系。BERT-LSTM和BERT-BiLSTM模型在整体性能上优于前3种模型，分类效果更佳，也体现出预训练模型较优的特征表征能力。其中，BERT-BiLSTM模型的性能最佳，因为BiLSTM结

构能更好地捕捉文本中的上下文信息，从而提高分类准确率。

4.3　不同信息源及其组合方式对分类结果的影响

论文信息源语料包括标题、关键词、摘要、全文内容等多个字段的信息，但从实现论文分类来讲，需要考虑成本、信息冗余性、信息可用性等方面，往往需要选择合适的信息源组合以产生合理的分类效果。将多个信息源字段对应的文本进行拼接，提取不同字段特征，进行分类对比，结果如表6所示。“标题+摘要+关键词”的信息源产生了最优的论文分类性能。

表6　不同信息源预测结果准确率

信息源	准确率
标题	0.869 7
关键词	0.642 6
摘要	0.749 7
标题+关键词	0.879 4
标题+摘要	0.882 8
摘要+关键词	0.799 3
标题+摘要+关键词	0.903 9

上述实验结果表明，标题、关键词、摘要及其不同组合方式均对论文分类的准确率有一定的影响。标题在论文分类中比较重要，基于其的分类准确率较高。摘要作为文本的总结和概述，在分类任务中也发挥了关键作用。虽然关键词大多来源于摘要，但在加入关键词字段特征后，模型分类性能仍然有所提升。通过结合标题、摘要和关键词等信息，模型能够更全面地理解论文内容，提高分类准确率。在论文分类任务中，充分考虑多个信息源的特征对于提升分类性能至关重要，这种综合利用多信息源的方法有助于模型更好地捕捉文本的特征，从而提高论文分类的准确率。

5　结论

本研究在融合BERT模型与BiLSTM模型的基础上，运用提示学习框架缩小预训练模型与下游任务之间的差距，将爬取到的论文标题、摘要和关键词融合，构建了学术论文自动分类系统。通过对论文文本进行预处理，剔除无关词语，选择关键分词进行特征关联，构建用于自动分类的文本分类系统。基于提示学习模板对比实验结果可知，在加入提示学习方法后，模型准确率提升了约3%，并且在实际应用中，基于提示学习的文本分类方法同时适用于少样本状况下的文本分类。该系统在有效提升论文分类的准确率和召回率的同时，最大限度地实现了学术资源分类的可视化和简洁性，加强了预训练模型与下游任务的联系，明显提升了信息服务效率和信息处理质量。在未来的研究中，考虑获取论文文本内容特征，更高效、快速地实现长距离依赖特征信息提取，以形成更精确、细致的论文自动分类系统。

参考文献

[1] 田亮，李博闻，章成志. 基于学术论文全文的跨语言研究方法自动分类研究[J]. 图书馆建设，2022（1）：75-86.

[2] SUN C，QIU X P，XU Y G，et al. How to fine-tune BERT for text classification? [M]//SUN C，QIU X P，XU Y G，et al. China National Conference on Chinese Computational Linguistics. Cham：Springer，2019：194-206.

[3] 李南星. 基于BERT和提示学习的改进句向量文本表示[D]. 汕头：汕头大学，2022.

[4] 余新言，曾诚，王乾，等. 基于知识增强和提示学习的小样本新闻主题分类方法[J/OL]. 计算机应用：1-10[2024-01-23]. http://kns.cnki.net/kcms/detail/51.1307.TP.20240111.1619.003.html.

[5] 岳增营，叶霞，刘睿珩. 基于语言模型的预训练技术研究综述[J]. 中文信息学报，2021，35（9）：15-29.

[6] 谢庆恒. 基于多源信息融合的学位论文自动分类标引[J]. 情报工程，2023，9（3）：70-80.

[7] 齐斌，邹红霞，王宇. 基于加权局部线性KNN的文本分类算法[J]. 计算机应用研究，2020，37（8）：2381-2385，2408.

[8] 程彭圣男，刘雪梅，李海瑞. 基于BERT-BiLSTM模型的输水工程巡检文本智能分类[J]. 中国农村水利水电，2023（10）：150-155，160.

[9] 穆建媛，朱毅，周鑫柯，等. 基于提示学习的中文短文本分类方法[J]. 中文信息学报，2023，37（7）：82-90.

[10] 章成志，李卓，储荷婷. 基于全文内容的学术论文研究方法自动分类研究[J]. 情报学报，2020，39（8）：852-862.

[11] 杨秀璋，夏换，于小民，等. 基于多视图融合的论文自动分类方法研究[J]. 现代电子技术，2020，43（8）：120-124.

[12] 王末，崔运鹏，陈丽，等. 基于深度学习的学术论文语步结构分

类方法研究[J]. 数据分析与知识发现，2020，4（6）：60-68.

[13] 李湘东，刘康，丁丛，等. 基于知网语义特征扩展的题名信息分类[J]. 图书馆杂志，2017，36（2）：11-19.

[14] 张智雄，赵旸，刘欢. 构建面向实际应用的科技文献自动分类引擎[J]. 中国图书馆学报，2022，48（4）：104-115.

[15] THIRUMOORTHY K，MUNEESWARAN K. Feature selection for text classification using machine learning approaches[J]. National Academy Science Letters，2022，45（1）：51-56.

[16] SADR H，SOLEIMANDARABI M N. ACNN-TL：attention-based convolutional neural network coupling with transfer learning and contextualized word representation for enhancing the performance of sentiment classification[J]. The Journal of Supercomputing，2022，78（7）：10149-10175.

[17] 李杰，李欢. 基于深度学习的短文本评论产品特征提取及情感分类研究[J]. 情报理论与实践，2018，41（2）：143-148.

作者简介

刘爱琴，女，博士，副教授，研究方向：信息技术与信息服务，E-mail：km_aql@sina.com。
贺玉斌，男，硕士研究生，研究方向：信息技术与信息服务。
马茹茹，女，硕士研究生，研究方向：信息技术与信息服务。

Automatic Classification of Academic Papers Based on Mixed Model of Prompt Learning

LIU AiQin HE YuBin MA RuRu
(School of Economics and Management, Shanxi University, Taiyuan 030006, P. R. China)

Abstract: The classification of academic papers is of great significance in knowledge management, academic exchange, research orientation, and academic evaluation. This paper builds an automatic classification system for academic papers based on a deep learning model. Compared with existing text classification methods, this system integrates the idea of prompt learning and better bridges the gap between the pre-training model and downstream tasks. The results show that this system can better improve text classification performance and standardization level, and provides better ways for researchers to manage, utilize, and mine information.

Keywords: Academic Paper; Prompt Learning; Automatic Classification

（责任编辑：王玮）