基于BERT-MLDFA的内容相近类目自动分类研究*
——以《中图法》E271和E712.51为例

李湘东[1,2] 石健[1] 孙倩茹[1] 贺超城[1]
(1. 武汉大学信息管理学院，武汉 430072；2. 武汉大学电子商务研究与发展中心，武汉 430072)

摘要：针对《中图法》中具有关联度大、区分度小等特点的内容相近类目，探讨利用深度学习来提升分类效果的方法。本文构建BERT-MLDFA模型，即通过多层级注意力机制对BERT不同层参数进行动态融合，并在任务数据集上预训练，进而以《中图法》中E271和E712.51作为典型内容相近类目进行自动分类实验。结果表明：本文方法的Macro_F1值达到0.987，相较于经典机器学习方法提升2.4%，而且该方法可以捕捉内容相近类目文本之间的细微语义差别，能够较好地应用于《中图法》以及其他内容相近类目分类，具有较强普适性。

关键词：《中图法》；深度学习；BERT；自动分类
中图分类号：G250.7　DOI：10.3772/j.issn.1673-2286.2022.02.003
引文格式：李湘东，石健，孙倩茹，等. 基于BERT-MLDFA的内容相近类目自动分类研究——以《中图法》E271和E712.51为例[J]. 数字图书馆论坛，2022 (2)：18-25.

基于科学分类体系的分类法在网络信息资源的系统组织和知识导航中具有重要作用。但随着数字资源的激增，人工分类组织效率低下导致其中一些资源网关已停止服务，迫切需要自动分类技术来解决上述问题[1]。在《中国图书馆分类法》（以下简称《中图法》）中，同一大类的众多下位类之间内容十分相近、语义关联度大、区分度小[2]，这些下位类在自动分类研究中称为内容相近类目，是人工分类和自动分类的共同难点。

《中图法》中存在大量内容相近类目，例如E大类下的E271（中国陆军）和E712.51（美国陆军）。这两个类别的书目信息虽然都是陆军主题，但所使用的词语里大多没有明确提及中国或美国等地区概念，难以在E大类的二级层次分类时予以区分至E2或者E7之中，更多是通过“游击队”和“野战排”、“坑道战中使用手榴弹”和“丛林战中使用手雷”等细微语义差别来区分，给基于机器学习方法的自动分类任务造成极大的困难。为切实提高区分两个内容相近类目文本之间细微语义差别的能力，本文以《中图法》中两个内容相近类目的书目信息作为分类对象，针对目前主流的BERT（Bidirectional Encoder Representations from Transformers）深度学习模型不能充分利用学习到的全部语义信息的缺点，通过多层级注意力机制对BERT不同层参数进行动态融合，提出改进的BERT-MLDFA（BERT with Multi-Layers Dynamic Fusion based on Attention）模型，并在该模型的基础上对任务数据集进一步预训练，提高分类效果。本研究提出的方法是实现3个及以上内容相近类目之间自动分类的核心技术，是实现《中图法》自动分类的基础性研究。

1 研究现状及意义

在《中图法》的自动分类研究中，分类的对象一般

* 本研究得到武汉大学青年研究中心调研课题“高校大学生‘内卷’机制的建模与仿真研究”（编号：20210407）资助。

是由题名、关键词和摘要等构成的文本信息，分类方法包括经典机器学习方法和深度学习方法。有学者[3-6]使用最近邻（K Nearest Neighbor，KNN）、朴素贝叶斯（Naive Bayes，NB）、支持向量机（Support Vector Machine，SVM）等经典机器学习分类算法中的一种或多种在《中图法》分类体系下对图书、网页或者其他类型的文献进行自动分类。近些年，基于长短期记忆模型（Long Short-Term Memory，LSTM）、卷积神经网络（Convolutional Neural Networks，CNN）和BERT等深度学习模型在自动分类领域取得了极大的成功，邓三鸿等[7]、郭利敏[8]和罗鹏程等[9]分别将LSTM、CNN和BERT用于《中图法》自动分类中，均取得了不错的分类效果。以上研究有一个共同点，即这些分类研究的对象一般是《中图法》上位类和中位类中比较容易区分的类别，没有聚焦于下位类中内容相近类目之间的难分类对象上。然而，《中图法》庞大的分类体系本身以及其特有的复分仿分机制使得同一大类下具有众多下位类，这些下位类别之间的主题非常接近、难以区分。因此，在自动分类时需要模型能够捕捉《中图法》内容相近类目文本之间的细微语义差别。在《中图法》内容相近类目分类方面，已有为数不多的相关研究是基于经典机器学习方法的，李湘东等[10]基于KNN、NB和SVM经典机器学习方法实现内容相近类目的分类；此外，还通过改进互信息特征选择法实现内容相近类目特征提取，并结合KNN分类算法实现内容相近类目的分类[11]，尚未见使用深度学习方法的相关研究。这些经典机器学习方法在处理文本时未考虑词语的上下文语义信息，而LSTM、CNN和BERT深度学习方法在一定程度上考虑了词语的上下文语义关系或者局部语义关系，在捕捉细微语义差别的能力上强于经典机器学习方法。因此，《中图法》内容相近类目自动分类需要探索使用深度学习方法，以取得更好的分类效果。

《中图法》中内容相近类目由2个及以上类目构成，需要二分类或者多分类技术对其进行自动分类。3个及以上类目的多分类问题可以通过一对一分解转换为多组二分类问题，因此二分类是多分类的基础[12]。目前，二分类技术主要集中在自动分类研究中的情感二分类上[11]，例如微博评论情感分析等。在两类择一的分类目标上以及两个类目的文本内容高度相似方面，情感二分类与《中图法》中两个内容相近类目的自动分类极为相似。实际上，李湘东等[10-11]就是针对《中图法》中两个内容相近类目的分类时使用了二分类技术。Li[13]和Ling[14]等指出微博情感分析实际上是一个将微博评论信息归类为积极或者消极的二分类问题，归类的难点在于评论信息中存在一些相似性极高却从属不同情感词，以及同一个词语在不同的语义环境下表达相反的情感，这些词语造成不同类目之间文本的高度相似性。现有研究[15-17]通过LSTM、CNN和BERT等深度学习模型获取这些词语在文中的语义信息，并应用于微博情感分析，取得了不错的分类效果，其中BERT表现最好。为了解释BERT为何能够取得很好的分类效果，Jawahar等[18]证明了具有12层级结构的BERT的不同层学习到了不同的语义信息，在BERT的底层、中间层和顶层分别学习到了表面特征、句法特征和语义特征，BERT利用顶层学习的语义特征信息为BERT的分类效果打下了良好的基础。但是BERT在做分类任务时，只利用最后一层参数进行分类，忽略了BERT其他层学习的语义信息。为了利用这些语义信息，李宁健等[19]通过CNN连接BERT的12层层级结构，提出BERT-MLF模型，并将该模型应用于情感分析任务中，取得了比BERT更好的分类效果。然而，BERT-MLF中的CNN结构不能为BERT不同层学习到的语义信息分配不同的权重，对BERT不同层参数进行动态融合时，在去除部分噪声语义信息的同时可能会丢失关键语义信息，从而导致分类性能下降。基于多层级注意力机制对BERT的12层参数赋予不同的权重是一个很好的思路，能为关键语义信息和噪声语义信息做自适应的权重分配，进而提升分类效果。本文使用的BERT是在使用中文维基百科等一般性语料上进行预训练所生成的，中文维基百科在内容上涵盖各学科领域以及社会生活的各个方面，具有较强的通用性，但也不能保证在面对文献分类等特定任务时具有较强的专指性。为此，需要针对具体的任务在上述中文维基百科等一般性语料的基础上进一步追加任务数据集继续进行预训练（Task-Adaptive Pretraining，以下简称“TAPT操作”）[20]。TAPT操作使BERT及其改进模型在具体任务上具有较强的专指性，通过扩大TAPT操作时任务数据集的内容使得模型在该任务上的专指性范围更广，或者通过更换TAPT操作的任务数据集使得模型在各自的任务上都具有各自相应的专指性，因此BERT及其改进模型结合TAPT操作可以适用于任务范围的扩大以及任务的更换，具有较强普适性。因此，针对《中图法》内容相近类目分类，BERT及其改进模型结合TAPT操作，分类效果可以得到更大程度的提升。

基于在《中图法》内容相近类目分类中缺乏并且需要深度学习方法的现状，同时内容相近类目分类研究主要集中在二分类方向上，本文在实验对象的选择上使用《中图法》两个内容相近类目开展二分类，并采用LSTM、CNN、BERT深度学习模型对其进行自动分类，比较这些深度学习方法相对于KNN、NB和SVM经典机器学习方法的优越性。针对BERT相比于LSTM、CNN分类效果更好并且BERT未能充分利用全部语义信息的现状，本文基于注意力机制对BERT不同层参数进行动态融合，提出改进的BERT-MLDFA模型。针对BERT及其改进模型结合TAPT操作可以更大程度提升分类效果，文本在BERT-MLDFA模型的基础上进行TAPT操作，以优化对《中图法》内容相近类目进行自动分类的效果。本文在《中图法》两个内容相近类目之间进行二分类研究，为实现3个或3个以上内容相近类目之间的自动分类打下更好的基础，具有较强的理论意义和实践价值。

2 基于深度学习的自动分类框架

2.1 基于LSTM和CNN的自动分类框架

LSTM和CNN是文本分类中两个基础且经典的深度学习模型，但在《中图法》内容相近类目自动分类中缺乏相关的应用研究。LSTM模型是一种特殊的循环神经网络结构，采用遗忘门、输入门和输出门3个门控函数获取文本序列的时序关系，从而取得文本特征之间的上下文信息。CNN主要由输入层、卷积层、池化层组成，卷积层通过卷积核获取特征之间的局部信息。LSTM忽略了特征之间的局部信息，CNN忽略了特征之间的上下文信息，因此二者在自动分类中各有优劣[21]。LSTM和CNN在自动分类时，通常结合Word2Vec词嵌入模型，以获取更好的分类效果，成为惯用的分类方法[22-23]。因此，针对《中图法》内容相近类目自动分类，本文首先采用典型的深度学习模型LSTM和CNN，结合Word2Vec词嵌入模型，设计相关研究框架，如图1所示。

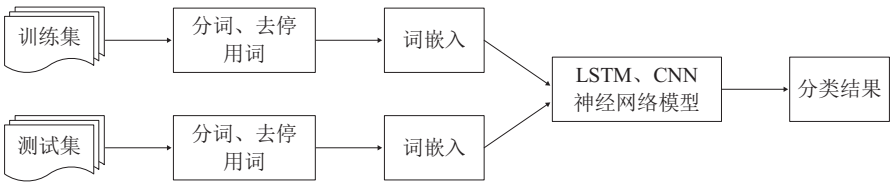

图1 基于LSTM和CNN的自动分类框架

基于LSTM和CNN的文本分类过程主要分为如下4个步骤。

（1）构建词语特征集合。具体包括，对训练集和测试集的文本使用python的jieba词库进行分词，并采用哈工大停用词表对分词结果去停用词，得到训练集和测试集的词语特征集合。

（2）词嵌入表示。具体包括，针对前述步骤（1）得到的训练集和测试集的词语特征集合，使用在维基百科语料上训练的Word2Vec词嵌入模型对其进行词嵌入表示，得到训练集和测试集的词嵌入特征表示集合。

（3）创建并训练模型。具体包括，创建LSTM和CNN分类模型，并将前述步骤（2）得到的训练集文本的词嵌入表示集合输入到神经网络模型中的词嵌入层作为LSTM和CNN的输入，然后对模型进行训练，得到训练好的LSTM和CNN神经网络模型。

（4）分类预测。具体包括，将前述步骤（2）得到的测试集文本的词嵌入表示集合输入到前述步骤（3）中已经训练好的LSTM和CNN神经网络模型中进行分类预测，得到分类结果。

2.2 基于BERT的自动分类框架

BERT是基于双向的Transformer模块结合而成的多层级结构[24]，在预训练过程中，采用遮罩语言模型（Masked Language Model，MLM）和下一句预测（Next Sentence Prediction，NSP）生成深度的双向语言表征，通过位置编码获取特征的上下文位置关系，从而根据上下文得到特征的动态向量表示，在自动分类上取得了比LSTM和CNN更好的效果，成为目前的主流模型。为了提升《中图法》内容相近类目自动分类效果，本文采用BERT模型并设计研究框架，如图2所示。

基于BERT模型的自动分类过程主要分为如下4个

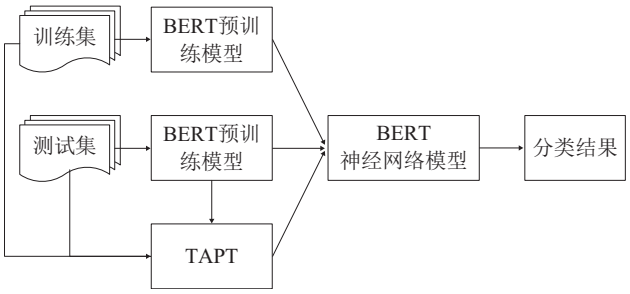

图2　基于BERT的自动分类框架

步骤。

（1）构建BERT特征向量表示。将训练集和测试集文本按照BERT预训练模型的输入格式进行预处理，构造特征向量表示，特征向量包括字向量、分段向量和位置编码向量3个部分。

（2）创建模型。创建BERT分类模型，并将BERT预训练模型权重初始化到BERT模型中。BERT模型结合TAPT操作，则将初始化权重之后的BERT模型进一步在任务数据集上预训练，预训练包括MLM任务和NSP任务[24]，并将权重更新到BERT模型中。

（3）训练模型。将前述步骤（1）得到的训练集BERT特征向量表示输入到前述步骤（2）中创建的BERT模型中进行训练，对BERT参数进行微调，得到训练好的BERT分类模型。

（4）分类预测。具体包括，将前述步骤（1）得到的测试集BERT特征向量表示输入到前述步骤（3）中训练好的BERT模型中进行分类预测，得到分类结果。

2.3　基于改进的BERT-MLDFA的自动分类框架

在做分类任务时，BERT只在最后一层参数上连接全连接层做分类，忽略了其他层学习的语义信息。为了进一步提升《中图法》内容相近类目的分类效果，本文对EBRT模型进行改进，提出一种改进的BERT-MLDFA模型，该方法基于注意力机制对BERT不同层特征进行融合，在融合过层中赋予不同层特征不同的权重并且权重在训练过程中自适应更新，从而可以充分利用BERT不同层学习的语义信息，得到语义信息丰富的特征表示，使得模型更好地学习和区分内容相近类目的文本类别。基于改进的BERT-MLDFA模型自动分类框架如图3所示。

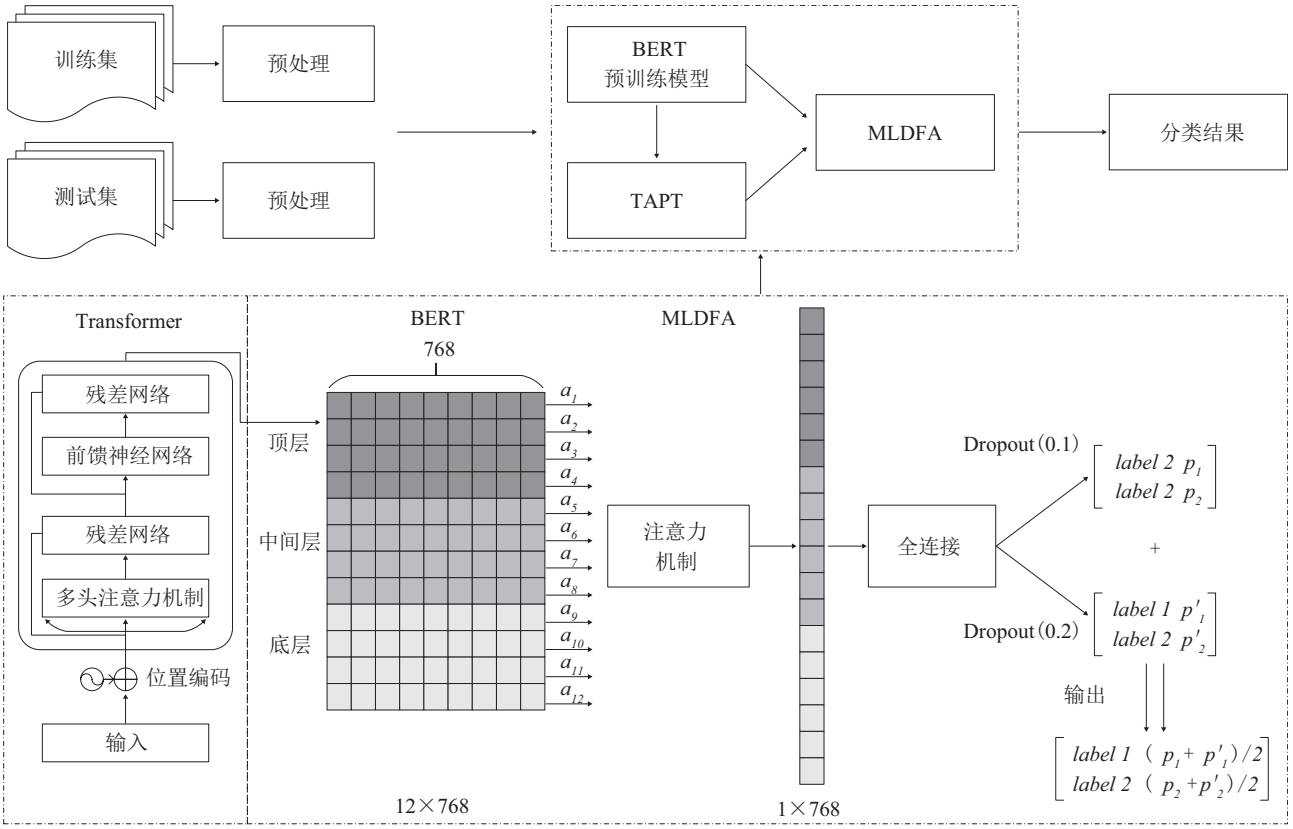

图3　基于BERT-MLDFA的自动分类框架

BERT-MLDFA模型首先将BERT的12层的每一层参数都经过最大池化计算，得到隐含状态h作为注意力机制层的输入，基于注意力机制为12层参数赋予不同的权重，得到融合特征*C*，融合特征*C*经过全连接层和softmax计算，通过两次Dropout取平均值作为最终分类概率，两次Dropout比率分别取值为0.1和0.2。

基于BERT-MLDFA模型的文本分类过程和BERT模型大致保持一致，需要将步骤（2）中创建的BERT模型替换为BERT-MLDFA模型，在使用BERT预训练模型对BERT模型进行初始化的同时，需要随机初始化BERT-MLDFA模型中的注意力机制层参数。

3 实验设计与分析

3.1 实验材料与评价方法

本文的实验对象是《中图法》中E271与E712.51两个类目，其原因在于这两个类目的典型性以及与过往研究在分类效果上的可比性。从文本用词方面看，这两个类同属军事主题，仅有中国和美国地区不同，文本用词相似，导致文本内容十分相近（见表1），因此，E271与E712.51能够比较好地代表使用自动方法难以区分的同一大类下的众多相似的下位类。从《中图法》体系结构上看，在二级类目上，E7的专类复分表与E2虽然并不完全一致（这也是未直接仿E2分的原因），但体系极其类似，具体到E271和E712.51两个类目，虽然在类目体系上不是复分仿分，但属于相同的主题，这与《中图法》对于地区与主题的复分仿分机制所产生的众多类目在内容的高度相似性上是一致的，在这两个类目上检验的自动分类方法可以有效地应用于其他复分仿分机制所产生的类目。从实验的严谨性上看，这两个类作为内容相近类目的分类对象，已有基于经典机器学习方法的相关成果，将其与本文的研究方法进行对比分析，以科学地验证深度学习方法相比于经典机器学习方法的优越性以及本文方法的有效性。

表1 E271与E712.51数据样本

类 目	文本样例
E271	SR-5型多口径多管火箭炮，该炮主要特点是射程远、威力大、自动化程度高、机动性好、操作简便
E712.51	M220式227毫米多管火箭炮，射程远、威力大、机动性好、反应速度快、自动化程度高、是目前世界上最先进的火箭炮

本文从维普数据库中提取E271和E712.51两个内容极为相近的语料作为实验的数据来源。其中，一共搜集E271的文档共616篇，E712.51文档1 366篇。每篇文档包括题名、关键词和摘要三部分信息，且两类文本数据集不存在交叉现象。对文本长度按照字符数进行统计，文本长度最短为37字符，最长为664字符，80%的文本长度集中在50~300字符之间。

以图书为例，即便拥有1 000万种（不含复本）图书的大型图书馆，在《中图法》5万多个类目中，平均每一个类目不到200册，因此，针对少样本的自动分类方法研究必须考虑今后应用时的可供实际使用的数据量。在实验对象的样本数量选择上，本文选择训练集的数量为200篇。为保证实验结果不受随机性和不平衡数据的影响，本文采用平衡数据集并将实验材料分为5组，每组实验材料在E271和E712.51中随机抽取200篇文档作为训练集，为保证训练集和测试集的文本无重复，在剩余的文档中随机抽取100篇文档作为测试集。分别对5组实验材料进行实验，记录每组实验结果，取5组实验结果的平均值作为最终实验结果。

为验证本文提出方法对内容相近类目分类的有效性，本文综合准确率和召回率计算F1值[25]，由于实验材料中两个类目的文本数量相等，宏平均F1值（Macro_F1）和微平均F1值（Micro_F1）保持一致，因此本文以Macro_F1值代表实验的分类效果，Macro_F1值越接近于1，分类效果越好。

3.2 实验环境及参数设置

本文实验基于Ubuntu操作系统，显存的大小为16G，并以Python编程语言和Torch1.8深度学习框架搭建神经网络结构。在预备实验中，确定了LSTM、CNN、BERT及其改进模型的超参数，包括学习率、批处理大小、训练迭代次数、文本最大长度等。LSTM和CNN的超参数取值分别为1e-3、60、30、512，BERT及其改进模型的超参数取值分别为2e-5、9、10、512。

3.3 实验结果与分析

对于内容相近类目的二分类，本文设置三组对比实验。首先，基于LSTM、CNN和BERT等深度学习方法对《中图法》的E271和E712.51进行自动分类，研究深度学习方法相比于KNN、NB和SVM等经典机器学

习方法的优越性；其次，本文基于注意力机制对BERT不同层参数进行动态融合，提出改进的BERT-MLDFA模型，并与基于CNN对BERT不同层参数进行融合的BERT-MLF模型进行对比，分析本文方法的优越性；最后，在BERT预训练模型的基础上，进一步在E271和E712.51的书目信息上预训练，对比分析TAPT操作的效果。

第一组对比实验的基准实验是基于KNN、NB和SVM等经典机器学习分类算法的分类效果，因此取文献[11]中在不同参数组合下的最优结果作为基准实验结果，并与本文采用的LSTM、CNN和BERT深度学习方法取得的实验结果进行比较。KNN、NB和SVM的Macro_F1值分别为0.951、0.959和0.963，LSTM、CNN和BERT的Macro_F1值分别为0.966、0.964和0.980。从实验结果可以看出，针对《中图法》内容相近类目分类，在KNN、NB、SVM经典机器学习方法中，SVM表现最好，相比于KNN和NB，Macro_F1值分别提升1.2%和0.4%；在LSTM、CNN、BERT深度学习方法中，BERT表现最好，相比于LSTM和CNN，Macro_F1值分别提升1.4%和1.6%；本文采用的3种深度学习方法整体优于3种经典机器学习方法，BERT相比于SVM的Macro_F1值提升1.7%。

第二组的对比实验是BERT、BERT-MLF模型与本文提出的BERT-MLDFA模型分类效果对比，3个模型的Macro_F1值分别为0.980、0.981和0.983。从实验结果可以看出，针对《中图法》内容相近类目分类，本文提出的BERT-MLDFA模型表现最好，相比于BERT提升0.3%，相比于BERT-MLF提升0.2%，在BERT的0.980的基线效果上更接近于1。

第三组对比实验是对比分析TAPT操作在BERT及其改进模型中的效果。BERT、BERT-MLF和BERT-MLDFA结合TAPT操作的Macro_F1值分别为0.983、0.983和0.987。从实验结果可以看出，BERT及其改进模型进行TAPT操作之后，Macro_F1值都有所提升，其中BERT-MLDFA结合TAPT提升最明显，相比于BERT-MLDFA提升0.4%，相比于BERT提升0.7%。同时，由于《中图法》中数目数量巨大，例如对于一个有1 000万种图书的大型图书馆，在《中图法》5万多个类目中，即使是0.7%的提升，也有可能使7万本图书被正确分类，能够带来巨大的时间和经济效益，因此具有很强的现实意义。

从以上三组对比实验分析可以得出以下3个结论。

（1）在《中图法》内容相近类目自动分类中，深度学习方法的分类效果优于经典机器学习方法。

（2）在《中图法》内容相近类目自动分类中，本文提出的改进的BERT-MLDFA模型表现最好，基于注意力机制对BERT不同层参数进行动态融合时可以结合文本的表面特征、句法特征、语义特征，能够捕捉关联度大、区分度小的文本之间的细微语义差别，证明了BERT-MLDFA在解决内容相近类目分类问题时的优越性。

（3）BERT及其改进模型在模型初始化权重之后，进行TAPT操作，即使用内容相似类目的E271和E712.51数据集继续进行预训练，可以进一步提升内容相近类目分类效果。针对其他内容相似类目进行分类，可以将E271和E712.51数据集更换为《中图法》上其他内容相似类目的数据集，因此，BERT及其改进模型结合TAPT操作具有较强普适性，可以有效应用于《中图法》以及其他内容相近类目分类中。

4 结语

针对内容相近类目的分类是《中图法》分类系统中一个十分重要的研究方向。由于内容相近类目文本之间关联度大、区分度小，在语义信息上只有细微差别，给自动分类带来了极大的困难。本文以《中图法》中E271和E712.51两个类别作为两个典型的内容相近类目，展开自动分类研究。实验结果表明，LSTM、CNN和BERT深度学习方法比KNN、NB、SVM更好；在深度学习方法中，BERT比LSTM和CNN更好；BERT-MLDFA模型能够获取内容相近类目文本之间的细微语义差别，相比于BERT分类效果进一步提升；BERT-MLDFA结合TAPT操作具有较强普适性，可以取得更好的分类效果。本文方法可以较好地应用于《中图法》以及内容相近类目分类中，但是BERT-MLDFA模型对BERT的不同层参数自适应的权重是如何分配的缺乏深入研究，导致该模型的可解释性不足是本文研究的一个局限。进一步探索本文提出的BERT-MLDFA模型的可解释性以及在其他更多内容相近类目中的应用，成为未来的研究重点。

参考文献

[1]　薛春香，何琳，侯汉清. 基于《中图法》知识库的自动分类相关

问题探析 [J]. 图书馆建设，2015（6）：16-20，26.

[2] 刘华梅. 《中国图书馆分类法》（第五版）类目复分仿分详解 [J]. 图书馆，2014（5）：128-131.

[3] WALTINGER U，MEHLER A，LÖSCH M，et al. Hierarchical classification of OAI metadata using the DDC taxonomy [M] // Advanced Language Technologies for Digital Libraries. Berlin：Springer，2009：29-40.

[4] PONG J Y H，KWOK R C W，LAU R Y K，et al. A comparative study of two automatic document classification methods in a library setting [J]. Journal of Information Science，2008，34（2）：213-230.

[5] WANG J. An extensive study on automated Dewey Decimal Classification [J]. Journal of the American Society for Information Science and Technology，2009，60（11）：2269-2286.

[6] 王昊，严明，苏新宁. 基于机器学习的中文书目自动分类研究 [J]. 中国图书馆学报，2010，36（6）：28-39.

[7] 邓三鸿，傅余洋子，王昊. 基于LSTM模型的中文图书多标签分类研究 [J]. 数据分析与知识发现，2017，1（7）：52-60.

[8] 郭利敏. 基于卷积神经网络的文献自动分类研究 [J]. 图书与情报，2017（6）：96-103.

[9] 罗鹏程，王一博，王继民. 基于深度预训练语言模型的文献学科自动分类研究 [J]. 情报学报，2020，39（10）：1046-1059.

[10] 李湘东，阮涛. 内容相近类目实现自动分类时相关分类技术的比较研究——以《中图法》E271和E712.51为例 [J]. 图书馆杂志，2018，37（6）：11-21，30.

[11] 李湘东，阮涛. 互信息特征选择法在《中图法》内容相似类目中的运用及改进——以E271和E712.51为例 [J]. 数字图书馆论坛，2018（1）：46-52.

[12] SANTHANAM V，MORARIU V I，HARWOOD D，et al. A non-parametric approach to extending generic binary classifiers for multi-classification [J]. Pattern Recognition，2016，58：149-158.

[13] LI J，FONG S，ZHUANG Y，et al. Hierarchical classification in text mining for sentiment analysis of online news [J]. Soft Computing，2016，20（9）：3411-3420.

[14] LING M，CHEN Q，SUN Q，et al. Hybrid neural network for Sina Weibo sentiment analysis [J]. IEEE Transactions on Computational Social Systems，2020，7（4）：983-990.

[15] SHI S，ZHAO M，GUAN J，et al. A hierarchical lstm model with multiple features for sentiment analysis of sina weibo texts [C] //2017 International Conference on Asian Language Processing（IALP）. IEEE，2017.

[16] YANG X，XU S，WU H，et al. Sentiment analysis of Weibo comment texts based on extended vocabulary and convolutional neural network [J]. Procedia Computer Science，2019，147：361-368.

[17] LI H，MA Y，MA Z，et al. Weibo text sentiment analysis based on BERT and Deep Learning [J]. Applied Sciences，2021，11（22）：10774.

[18] JAWAHAR G，SAGOT B，SEDDAH D. What does BERT learn about the structure of language? [C] //ACL 2019-57th Annual Meeting of the Association for Computational Linguistics. 2019.

[19] 李宁健，方睿. 融合BERT多层特征的方面级情感分析 [J]. 计算机科学与应用，2020，10（12）：2147-2158.

[20] GURURANGAN S，MARASOVIĆ A，SWAYAMDIPTA S，et al. Don't stop pretraining：adapt language models to domains and tasks [EB/OL]. [2022-01-01]. https://aclanthology.org/2020.acl-main.740.pdf.

[21] 黄金杰，蔺江全，何勇军，等. 局部语义与上下文关系的中文短文本分类算法 [J]. 计算机工程与应用，2021，57（6）：94-100.

[22] XIAO L，WANG G，ZUO Y. Research on patent text classification based on Word2Vec and LSTM [C] //2018 11th International Symposium on Computational Intelligence and Design（ISCID）. IEEE，2018.

[23] SHARMA A K，CCHAURASIA S，SRIVASTAVA D K. Sentimental short sentences classification by using CNN deep learning model with fine tuned Word2Vec [J]. Procedia Computer Science，2020，167：1139-1147.

[24] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-01-01]. https://aclanthology.org/N19-1423/.

[25] SOKOLOVA M，LAPALME G. A systematic analysis of performance measures for classification tasks [J]. Information processing & management，2009，45（4）：427-437.

作者简介

李湘东，男，1963年生，博士，副教授，研究方向：文本分类、信息检索、数据挖掘，E-mail：xli_xiao@hotmail.com。
石健，男，1996年生，硕士研究生，研究方向：文本分类、数据挖掘。
孙倩茹，女，1997年生，硕士研究生，研究方向：文本分类、数据挖掘。
贺超城，男，1993年生，博士，讲师，研究方向：文本分类、数据挖掘。

Automatic Classification Research of Similar Categories Based on BERT-MLDFA: Take E271 and E712.51 in CLC as an Example

LI XiangDong[1,2] SHI Jian[1] SUN QianRu[1] HE ChaoCheng[1]
(1. School of Information Management, Wuhan University, Wuhan 430072, P. R. China; 2. Center for Electronic Commerce Research and Development at Wuhan University, Wuhan 430072, P. R. China)

Abstract: This paper discusses the method of using deep learning method to improve the classification performance of the similar categories in *Chinese Library Classification* which have the features of high correlation degree and low differentiation degree. This paper proposes a BERT-MLDFA model that dynamically integrates parameters of different BERT layers through multi-level attention mechanism, and further pretrains on task datasets. Then, to conduct automated classification experiments, E271 and E712.51 in *Chinese Library Classification* were used as typical similar categories. The results show that the Macro_F1 value of the proposed method reaches 0.987, which is 2.4% higher than that of the classical machine learning method. The method proposed in this paper can capture the subtle semantic differences between texts of similar categories, which can be applied to *Chinese Library Classification* and other similar categories and is universal.

Keywords: *Chinese Library Classification*; Deep Learning; BERT; Automatic Classification

（收稿日期：2022-02-03）

征文通知

2022年中国科学技术信息研究所博士后科研工作站设站20周年学术&主题征文通知

2002年中国科学技术信息研究所获批成为国内首家具有独立招收资格的“图书情报与档案管理”学科博士后科研工作站。2022年，20周年之际，为回顾和总结图情档学科博士后科研工作站的发展历程和建设成就，中信所拟汇聚学界同人就图情档学科发展、人才队伍培养、支撑科技决策实践等内容开展征文探讨。

本次学术&主题征文均组织专家评选，设立一、二、三等奖若干，颁发荣誉证书并给予一定奖励。获奖稿件将推荐到《中国软科学》《情报学报》《情报工程》《中国科技资源导刊》《数字图书馆论坛》《全球科技经济瞭望》《高技术通讯》等期刊遴选发表。获奖作品将集结成册，收录到设站20周年纪念册。

一、学术征文选题

（1）新时代图情档学科发展研究。

（2）图情档学科博士后“学—研—产”实践创新研究。

（3）图情档学科设站单位高水平人才培养特色研究。

二、主题征文选题

（1）表达对工作站设站20周年的真情实感等。

（2）回忆或记录在中信所从事博士后科研工作的心路历程和经历故事等。

（3）介绍作为博士后合作导师的指导经验方法与心得体悟等。

（4）回顾设站20年的发展历程，以见证者、亲历者的身份记述工作站的发展变化，讲述20年来各个发展时期的重大事件、重要人物、发展成果等。

三、征文要求

（1）投稿文章必须是未公开发表的原创性研究成果。学术征文要求论据充分、数据可靠、结构严谨，字数在6 000字以上，格式参考《数字图书馆论坛》投稿模板。主题征文体裁不限，字数在2 000字以上（诗歌除外）。

（2）投稿请发送至huayf@istic.ac.cn（邮件主题注明“设站20周年学术征文/设站20周年主题征文”，论文以“文章名称—作者姓名”的方式命名，稿件内附个人简介，包括作者姓名、联系方式、工作单位、职务等基本信息）。

（3）投稿截止日期为2022年6月30日。

联系人：花老师

联系方式：010-58882046