基于BERT模型的图书表示学习与多标签分类研究*

蒋彦廷 胡韧奋

摘 要 中文图书细粒度多标签分类的自动化，有利于促进图书的检索与学科的沟通。文章充分发挥BERT语言模型的微调特性，提出一种通过21类粗粒度分类微调语言模型，学习到更好的图书表示，进而实现细粒度分类的新策略。结果显示，在单标签的分类任务上，BERT模型的正确率分别较LSTM与Fasttext模型提升约4.9%与2.0%。KNN-ML对257类的细粒度多标签分类证明了前期微调的有效性。最佳情况下，有75.82%的图书细粒度类别恰好全部预测正确，92.10%的图书至少被正确预测了一个细粒度类别。因此可以得出结论，该系统有助于实现图书自动的细粒度归类，并帮助图书标引者补充合理的分类号。

关键词 中文图书 BERT模型 深度学习 微调策略 多标签分类
分类号 G254.1
DOI 10.16810/j.cnki.1672-514X.2020.09.007

Representation Learning and Multi-label Classification of Books Based on BERT

Jiang Yanting, Hu Renfen

Abstract The automation of the fine-grained multi-label classification of Chinese books is beneficial to the book index and subject communication. This paper makes full use of fine-tuning of BERT model and puts forward a novel strategy which fine-tunes the model on the coarse-grained classification task to learn a better book representation, and then completes the multi-label classification. The result shows that on the single-label classification, the accuracy of BERT has increased by about 4.9% and 2.0% compared with LSTM and Fasttext. The classification result of KNN-ML indicates the effectiveness of fine-tuning. Under the best situation, 75.82% of books are correctly sorted out, and 92.10% of books are predicted with at least one correct label. It draws a conclusion that this system is of great benefit to automatic fine-grained classification, and can help book annotators replenish the potential missing category code.

Keywords Chinese books. BERT. Deep learning. Fine tuning. Multi-label classification.

中国近年来的图书出版规模十分可观。据统计，2016至2018年国内年均申报各类图书选题29.5万余种[1]。伴随各学科的发展与相互交融，越来越多跨学科、边缘学科、复合视野的研究成果以图书的形式呈现出来。这意味着用《中国图书馆分类法》（以下简称《中图法》）中的单一类别标签，已难以全面、准确地概括它们的主题与内容。然而囿于有限的精力与知识面，人工编制的图书在版编目（Cataloguing In Publication, CIP）给大多数图书只指定了1个分类标引，这在一定程度上限制了图书的检索与学科间的交流。因此，如何利用信息技术，自动补全原有图书可能缺失的分类号，并实现新图书自动的、细粒度归类，打通各专业学科之间的屏障，是图书情报领域值得研究的课题。图书自动分类是文本分类（Text Classification）的一个子领域，与其他类型的文本相比，学界对中文图书分类的研究相对较少。本文拟尝试一种基于BERT语言模型的模型的图书的粒度分类引法来解决自动分类中存在的问题。

1 相关研究回顾

在以往的成果中，中文图书分类的方法主要分为两种。一是基于特征工程的经典机器学习方法，二是自动编码提取特征的深度学习方法。前者如王昊等[2]在特征加权的基础上，采用支持向量机（SVM），构建了一个浅层的中文图书分类模型；刘高军等[3]、潘辉[4]混合采用TF-IDF、隐含狄利克雷分布（LDA）主题模型抽取图书特征，采用极限学习机算法实现图书分类。后者以邓三鸿、傅余洋子[5]等的研究为代表，基于

*本文系国家社科基金青年项目“面向汉语国际教育的智能测试技术研究”（项目编号：18CYY029）研究成果之一。

字嵌入与LSTM模型，通过构造多个二元分类器，对5类图书进行多标签分类实验。总的来看，目前的研究还存在提升的空间。第一，实验数据集涉及的类别较少，未反映出《中图法》的基本面貌。第二，总体上缺乏对图书多标签分类的关注，既有的图书多标签分类方法存在计算开销大、类别不均衡的问题。第三，图书分类号的精细程度与分类器的性能难以兼得。《中图法》是一个树状的、多层次的图书分类体系，如果只将一级大类作为分类标签，分类器无法预测更加具体的分类号；如果采用层次化的细粒度分类，则会存在类别过多、数据稀疏等问题[6]，且难以顾及兼类的图书。因此，如何细粒度地、准确地预测图书的分类号，是亟待探研的问题。

近年来，自然语言处理界以ELMo[7]、BERT[8]为代表的深度预训练语言模型（Pre-trained Language Model）极大改善了文本语义表示的效果，并在文本分类等各项下游任务中取得了明显突破。预训练语言模型应用于下游任务，主要分为两种策略：一是基于特征的（Feature-based）策略，将固定的语言特征向量从模型中提取出来，以ELMo为代表；二是微调（Fine-tuning）策略，一方面，模型顶部接入着眼于具体任务的分类层，另一方面，语言模型所有的参数也随着下游任务的训练适度优化，以BERT为代表[8]。由此，我们尝试提出一种基于BERT语言模型的图书细粒度分类的方法，首先尊重并充分利用原有的图书信息及分类标签，通过进一步预训练（Further Pre-training）与粗粒度的分类任务，让BERT模型微调经由图书向量相似度计算，实现图书的细粒度分类。

2　BERT模型介绍

BERT (Bidirectional Encoder Representations from Transformers)是一种基于Transformer架构的深度预训练语言模型，其结构主要如图1所示。

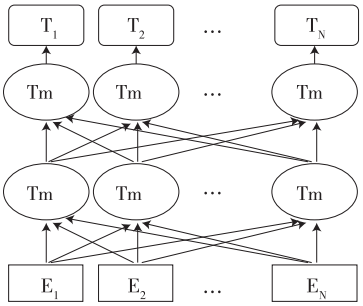

图1　BERT模型基本结构

以中文预训练模型为例，图1的E1,E2,…EN表示在首尾分别添加[CLS]和[SEP]标记的文本字符。它们依次经过12层双向的Transformer（Trm）编码器，就可以得到文本字符语境化的向量表示（Contextual Embeddings）。Transformer是一个基于自注意力（Self-attention）机制的编码-解码器。最底层的Transformer编码器的输入为字符向量、字符位置向量与句子片段向量之和。模型内每一层均由多头自注意力(Multi-head Self-attention) 和前馈神经网络（Feed-forward Neural Networks）两部分构成，前者使编码器在给每个字符编码时，能关注到周围其他字符的信息；后者用于增强模型的拟合能力。模型的每一层经过一个相加与归一化（Add & Norm）操作后，生成新的字符向量，作为下一层编码器的输入。顶层编码器输出的[CLS]标记的编码向量T1，可以视为整个句子的语义表征，用于后续的文本分类任务[9]。

另外，为增强语义表示的能力，BERT提出了遮罩语言模型（Masked LM, MLM）和下句预测（Next Sentence Prediction, NSP）的概念。MLM实质是一个完型填空任务，中文语料中15%的字会被选中，其中的80%被替换为[MASK]，10%被随机替换为另一个字，剩下的10%保持原字。模型需要经由一个线性分类器，预测被选中的字。出于与后面任务保持一致的考虑，BERT需按一定的比例在预测的字的位置放置原字或者某个随机字，使得模型更偏向于利用上下文信息预测被选中字。在下句预测任务中，模型选择若干句子对，其中有50%的概率两句相邻，50%的概率两句不相邻。模型通过上述两个目标任务，能够较好地学习到字词和句间的语义信息。

3　基于表示学习的图书粗粒度分类

我们尝试在图书粗粒度分类任务上对模型进行微调（Fine-tuning），提升预训练模型对图书数据表示的准确度，为后续的细粒度分类任务奠定基础。首先进行单标签分类，以测试BERT图书分类的有效字段，检验进一步预训练的效果，并与其他模型进行比较；既而进行多标签实验并讨论其实用性。本文的整体模型架构如图2所示。

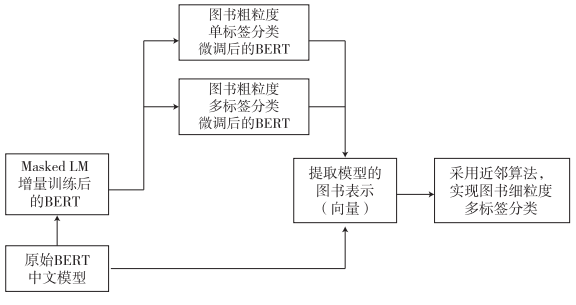

图2　基于BERT表示学习的图书分类模型架构

3.1 数据集与实验环境

图书数据[①]广泛采集自读秀学术网站。考虑到Z类（综合性图书）主要包括辞典、类书、年鉴等类型，出版数量较少，字段缺失的情况较多，我们采集了A–X共21大类、132 803册图书的书名、主题词、摘要、中图分类号等字段。在这些图书中，只有1个分类号的书为128 548册，占比约96.8%；拥有2个分类号的书达4152册，拥有3个及以上分类号的书为103册。

实验的操作环境为Ubuntu16.04.2LTS(GNU/Linux 4.8.0–36–genericx86_64)，采用2块1080ti型号的GPU，预训练语言模型为BERT基础（BERT–base–Chinese）版[②]，为12层的Transformer模型，hidden size为768，自注意力机制的head数量为12，总参数量为110M。

3.2 单标签分类实验

我们首先对只有1个分类号的图书进行实验。具体到各类别的图书数量如表1所示。

表1 单标签图书分类数据[③]

分类号	A	B	C	D	E	F	G	H
图书数量	5129	8052	6079	7209	4988	7955	10764	3882

分类号	I	J	K	N	O	P	Q	R
图书数量	4072	6824	9293	4424	4223	5821	6494	7680

分类号	S	T	U	V	X
图书数量	6859	6724	4532	2824	4720

对于单标签文本分类任务，BERT模型提取顶层的符号[CLS]的特征向量v（768维）作为整个文本的特征表示，再后接一个768*n的全连接层（Fully–connected layer）W（n为类别数量），最后通过softmax函数归一化，输出一个文本分别属于各个类别的概率：

$$P(c/v) = softmax(W \cdot v)$$

其中softmax函数：$softmax(x_c) = \frac{exp(x_c)}{\sum_{i=1}^{n} exp(x_i)}$

在训练过程中，模型会调整全连接层W以及BERT模型的参数，使得正确标签所对应的概率最大化。

在训练策略方面，我们将图书数据集的顺序随机打乱，按8:1:1的比例划分训练集、验证集和测试集，并参考Sun（2019）等人[10]在BERT上的分类实验经验，如下设置超参数：学习率lr=2e–5，衰变因子ξ=0.95。此外，训练遵循早停（Early stopping）原则，当模型的损失在验证集上不再下降，就视为模型在验证集上已经收敛，可以停止训练。

如图3所示，当书名与主题词字段作为分类字符串时，分类正确率比单一的书名大幅提升近8%。而在此基础上加入出版社名、摘要等字段，分类正确率上升幅度不明显，训练收敛需要的迭代次数却逐渐增多。综合考虑性能、训练次数与字段的常见性，我们认为“书名+主题词”能够扼要地表示图书的主要内容，将它们作为后续实验所用的字段，将对应的分类实验记为BERT–base–Chinese，作为后续实验的参考。

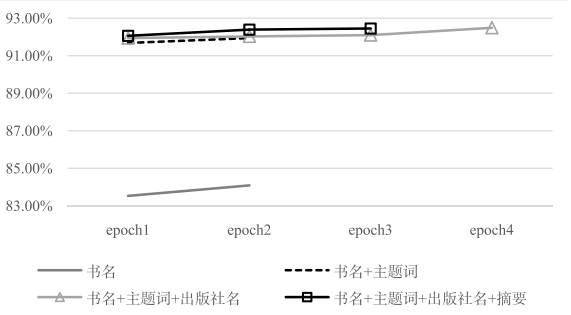

图3 BERT单标签图书分类在验证集上的正确率随迭代次数的变化

在训练基本版BERT中文模型时，Devin等[8]采用了字粒度的中文维基百科作为训练语料。Sun[10]等人检验了进一步预训练（Further pre–train）该语言模型对文本分类的有效性。我们尝试利用图书数据集增量训练语言模型。考虑到每一本书均表示为一个书名加若干主题词、按字切分的短文本，我们只采用如前文所述的遮罩语言模型（Masked LM）的训练策略，选择语料中15%的字进行预测，一共训练5900步，得到增量训练后的语言模型。在此基础上再进行21类图书的单标签分类实验，记作BERT–Increse。另外，本文将邓三鸿、傅余洋子等[5]提出的基于单向长短期记忆网络（LSTM）的图书分类模型，以及基于Facebook的Fasttext文本分类模型[④]作为基线（Baseline）模型。LSTM设置1层隐层，每个隐层含128个节点，并采用Adam优化算法[11]与早停策略。Baseline与BERT–base–Chinese、BERT–Increse实验均使用同样比例与内容的训练、验证、测试数据。

① 图书数据集地址：https://github.com/JiangYanting/Chinese_book_dataset

② https://s3.amazonaws.com/models.huggingface.co/bert/bert–base–chinese.tar.gz

③ A:马克思主义、列宁主义、毛泽东思想、邓小平理论；B:哲学、宗教、心理；C:社科总论；D:政治、法律；E:军事；F:经济；G:文化、科学、体育、教育；H:语言文字；I:文学；J:艺术；K:历史地理；N:自科总论；O:数理科学与化学；P:天文、地球科学；Q:生物科学；R:医药、卫生；S:农业科学；T:工业技术；U:交通运输；V:航空航天；X:环境科学、安全科学。

④ https://github.com/facebookresearch/fastText

如图4所示，进一步预训练的语言模型BERT-Incerse较BERT-base-Chinese能再获得约0.23%的正确率提升，表明通过遮罩语言模型（Masked LM）增量训练BERT对于文本分类也具有一定的功效。BERT-Increse模型在验证集上的正确率分别高出LSTM和Fasttext模型约4.9%与2.0%，并且前者需要训练的周期数比后者更少，这证明了我们基于BERT的图书分类方法的有效性。

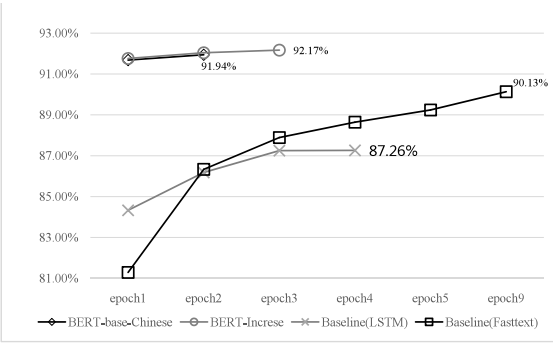

图4　单标签图书分类实验在验证集上的正确率对比

此外，如图5所示，在具体图书类别的F1值精度方面，A（马列主义等）、J（艺术）和U（交通运输）三类图书具有最佳的分类表现，而T（工业技术）、K（历史、地理）和N（自然科学总论）三类图书的分类F1值较低。这表明A、J、U类图书至少在书名、关键词上的分布较为集中。而T、K、N类图书涉及的子领域较多，话题更为广阔，数据相对稀疏。它们是人们进一步优化分类模型时，需要着重关注的对象。

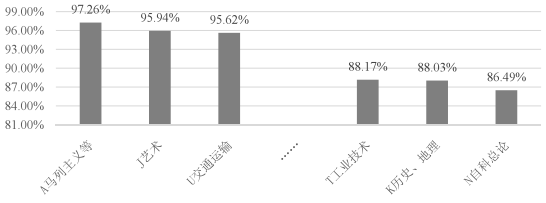

图5　BERT分类模型在验证集上具体图书类别的分类F1分数

3.3　多标签分类实验

在单标签分类的基础上，我们进行多标签的分类实验。除了前一小节所用的单标签数据外，又增加了4152册兼类的图书数据。虽然多标签的图书占数据集图书总数的比例较小（约3.2%），但图书兼类情况错综复杂，种类高达181种。兼类最多的两类情况是F兼D（经济类兼政法类）、R兼Q（医药卫生类兼生物科学类），图书分别达到124、105例。而兼类情况较少的如N兼I（自然科学总论类兼文学类）都仅有1例。这意味着在后续的多标签分类时，不宜简单地将兼类的图书单独划类，否则将面临类别过多、数据稀疏的问题。

多标签分类是文本自动分类中的一个研究热点与难点，其目的在于给测试集中每一个文本预测一个或多个可能的类别。根据假设的不同，以往的研究主要分为两大类。第一，假设各类别相互独立，不考虑类别之间的相关性，进而运用3种具体的分类算法：（1）二元相关（Binary Relevance）算法[12]，即把多标签分类转化为多个二类分类任务。（2）基于K近邻（KNN）改进的惰性学习算法[13]。（3）调整损失函数，独立地计算、输出一个文本属于各类的概率[14]。第二，考虑标签两两之间的相关性，由此设计出排序支持向量机（Rank SVM）[15]、双层的主题模型[16]等方法，以及基于深度学习序列生成的多标签分类方法[17]。具体到本任务，由于图书不存在诸如“属于甲类就一定属于乙类”或“属于丙类就一定不属于丁类”的情况，因此我们仍假设21个图书大类相互独立，将损失函数调整为带有sigmoid函数的二元交叉熵损失函数（Binary Cross Entropy with Logits Loss, BCE with Logits Loss），其中一个样本的损失如下式计算[14]：

$$\mathrm{BCE\_With\_Logits\_Loss}(x_n,y_n) = -\frac{1}{n}\sum[y_n \cdot ln\sigma(x_n) + (1-y_n)\cdot \ln(1-\sigma(x_n))]$$

其中sigmoid函数$\sigma(x_n) = \frac{1}{e^{(-x_n)}+1}$，$\sigma(x_n) \in (0,1)$

n为类别总数，x_n是模型的输出值，表示预测样本属于某一类别的概率；y_n是样本在某一类别下的真实标签，1表示属于该类别，0表示不属于该类别，是x_n的优化目标。与单标签分类通常采用的softmax交叉熵损失函数不同的是，sigmoid函数使一个样本属于各类别的概率分布在(0,1)之间，且没有进行类别之间的归一化，使各类别的概率之和可能大于1。这允许模型给每一个标签分配独立的概率。在测试阶段，模型将凡是概率大于50%的标签输出，作为一个样本多标签分类的预测结果。

我们基于上文BERT-Increse实验的模型进行多标签分类的微调。整个数据集按约8:1:1的比例划分训练集、验证集、测试集，一共训练2个epoch使模型在验证集上的损失收敛。模型在测试集13334个样本上预测的结果如表2所示。

表2　多标签分类在测试集上的预测结果

(1)：正确预测出至少一个分类号的样本比例	92.53%
(2)：正确预测出全部分类号的样本比例	89.98%
(3)：在(2)中多预测了分类号的样本比例	1.24%

值得注意的是，在符合情况（3）的165例样本里，一些预测虽然比实际标签数更多，但经人工检查发现，这些与实际标签不一致的预测也有其合理性，部分例子如表3所示。

表 3　人工标注与机器预测的图书大类对比

书名	人工标注的大类	机器预测的大类
毛泽东思想与中国铁路建设	F 经济	F 经济、A 毛泽东思想等
形式语言与自动机第 2 版	T 工业技术	T 工业技术、H 语言文字
当代世界经济与政治	F 经济	D 政治法律、F 经济
佛学与六朝文论	I 文学	I 文学、B 哲学 / 宗教
世界山川地理简介	K 历史地理	K 历史地理、P 地球科学等
生理学实验指导	Q 生物科学	Q 生物科学、R 医药卫生
中国沙区生态重建与恢复	P 地球科学等	P 地球科学等、X 环境安全

从表3中不难发现，机器多预测出的一些分类号其实无可厚非。例如《形式语言与自动机第2版》，其中论述的形式文法和自动机，既是程序语言编译技术的重要理论基础，可归为工业技术类下辖的TP类（计算机、自动化技术）；又是形式语言学、转换生成语法等语言学流派的研究内容，也可归为H语言文字类下辖的H087（数理语言学）类，宜按互见分类处理。经过人工检查统计，这165例样本中，至少有81%的预测有一定的合理性。这一方面表明基于BERT的粗粒度、多标签分类已具有一定的实用性，可以初步预测图书所属的学科大类，并能够补充一些图书可能缺失的分类号，为图书标引的工作者提供有益的推荐与参考；另一方面该分类任务也促使BERT通过微调（Fine-tuning）学习到更好的图书表示，为后续的细粒度分类打下了基础。

4　基于微调 BERT 与多标签 K 近邻的图书细粒度分类实现

为验证粗粒度分类任务对BERT模型图书表示的影响，我们尝试从不同阶段的模型中提取768维的图书向量，并通过主成分分析（Primary Component Analysis, PCA）技术降维可视化，分析图书分布的变化。我们以图书馆学情报学（属于G3与G2）、语言学（属于H0）与计算机自动化技术（属于TP）3类各100册图书为例，选取其书名、主题词字段作为输入词，观察它们的语义表示变迁，见图6至图9。

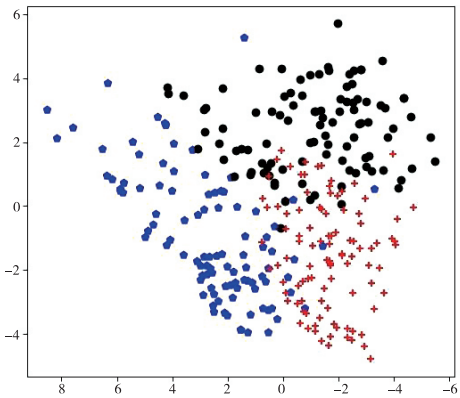

图 6　基于原始中文 BERT 的图书表示

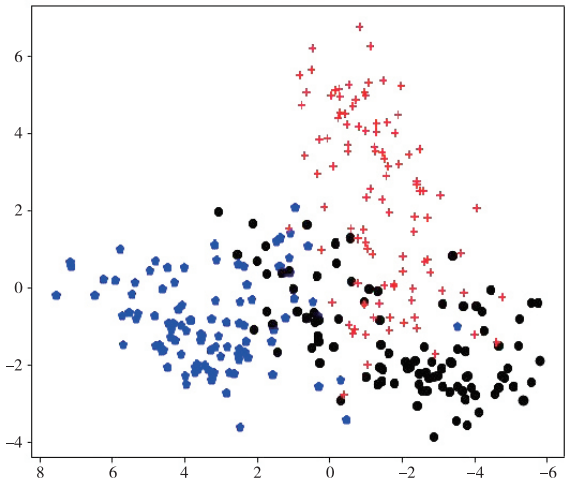

图 7 Masked LM 增量训练后的图书表示

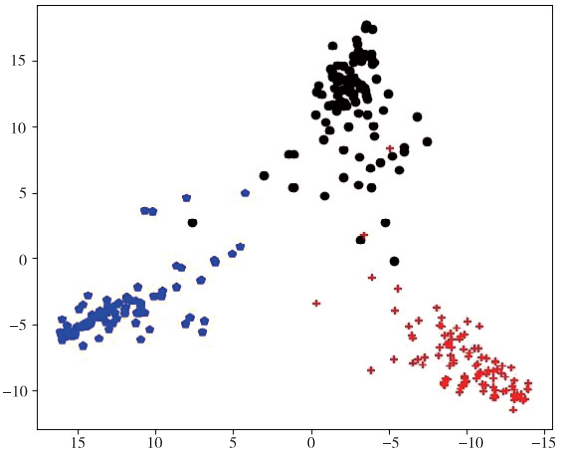

图 8　单标签分类 fine-tuning 后的图书表示

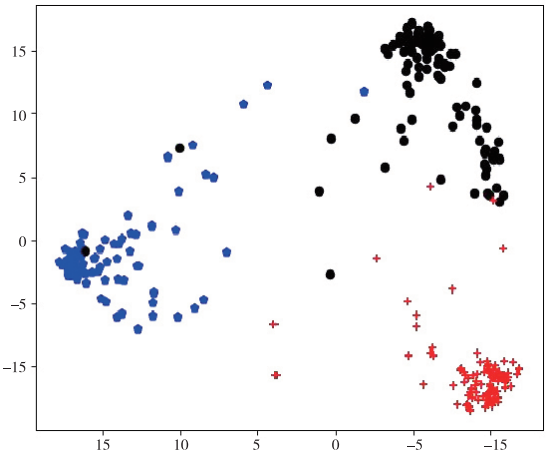

图 9　多标签分类 fine-tuning 后的图书表示[①]

综合图6 ~ 图9可以发现，从原始模型到增量训练，再到单标签分类、多标签分类后，三类图书的分布呈现出同类图书集聚、类间图书距离拉大的趋势。而该趋势在两个分类任务后的模型上表现得尤为明显。

[①]红色“+”表示计算机、自动化类图书；圆点表示语言学类图书；五边形表示图书馆学情报学类图书。

这反映出通过BERT在完成下游任务的同时，语言模型本身也发生了显著变化，通过编码图书主题和粗粒度类别的信息，图书语义表示比原始模型更加精准。

细粒度的多标签分类以数据集图书所属的中图法各二级学科作为类别，总计257类。训练集为3.3节粗粒度多分类实验中的训练集与验证集之和；测试集则保持不变。我们首先从粗粒度多标签分类微调后的BERT模型中提取768维的图书向量，然后采用基于K近邻（K-Nearest Neighbor, KNN）的分类方法KNN-ML（KNN-Multi Label）。具体过程如下：

（1）对于测试集中的每一个样本点y_i，查找出训练集中与之向量距离最近的k个近邻样本x1,x2,…,xk，记作集合U。

（2）统计集合U中，各样本所属的中图法二级学科类别C_i及其出现频次F_i。

（3）设置阈值L，当F_i >=L时，就将对应的C_i判定为测试样本点y_i所属的二级学科。

基于KNN-ML的多标签分类具有诸多优势，首先，它是一种惰性学习方法，节省了显性的训练过程；其次，KNN-ML仅以近邻范围内样本投票表决的方式进行预测，既实现了多标签分类，也无须计算全部257类的概率；第三，图书的向量表示直接从BERT模型中继承而来，增量训练、系统的后期维护也较为方便。

yi查找近邻样本时，KNN-ML利用球树（Ball Tree）的数据结构来优化查找的过程。球树使用超球面对样本空间进行划分，在查询一个测试样本的k近邻时避免了蛮力计算；此外，球树运用球心与半径描述样本点，样本点占用的空间对维数不敏感，这弥补了矩形划分在高维时存储较大、查询较慢的不足[18]。因此球树适用于本次实验向量维数较高的情况。

我们记原始的中文BERT模型为BERT-base-Chinese，在此基础上，记进行粗粒度单标签分类微调后的模型为BERT-single-clf，记粗粒度多标签分类微调后的模型为BERT-multi-clf。经多次参数调优，KNN-ML设置近邻数目k=10，观察各模型在阈值L变化时的表现变化。如表4所示。

表4　基于BERT与KNN-ML的图书细粒度多标签分类结果

阈值L	提取图书向量的模型类型	指标1：正确预测至少一个标签的样本占比	指标2：恰好预测出全部标签的样本占比	指标3：多预测了标签的样本占比
2	BERT-base-Chinese	91.43%	52.88%	36.12%
	BERT-single-clf	92.10%	56.20%	33.25%
	BERT-multi-clf	91.63%	61.42%	27.80%
3	BERT-base-Chinese	86.55%	67.38%	16.42%
	BERT-single-clf	88.31%	68.01%	17.38%
	BERT-multi-clf	88.10%	70.85%	14.64%
4	BERT-base-Chinese	81.54%	70.01%	5.54%
	BERT-single-clf	84.01%	74.23%	6.80%
	BERT-multi-clf	84.44%	75.82%	5.94%

根据表4，首先，未经微调的BERT-base-Chinese在指标1、2上的表现均不及微调后的两个模型。这证明BERT通过微调融入中图法一级大类的信息后，也能提高二级类别分类的准确度。其次，BERT-single-clf模型的总体效果居于第二，在不同阈值L下的9项指标评测中，有2项取得最佳。由于它仅在单标签分类任务上微调，缺乏对兼类图书的语义编码，因此BERT-single-clf至少正确预测出一个标签的能力较强（指标1），但其准确预测出全部标签的能力（指标2）略逊于BERT-multi-clf，且容易将非兼类的图书预测为兼类（指标3）。最后，综合3个指标，BERT-multi-clf模型的表现最佳，在9项指标中取得了6项最佳。当阈值L=4时，测试集中84.44%的图书被正确预测了至少1个二级类别；有75.82%的图书的分类号完全预测正确。这对于多达257个类别的细粒度多标签分类任务来说，依然是良好的表现，且性能明显优于原始BERT模型与单标签粗粒度分类后的BERT-single-clf。在占比5.94%的多预测了标签的图书中，部分例子如表5所示。

表5　部分图书人工标注与KNN-ML预测的二级类别对比

书名	人工标注的二级类别	KNN-ML预测的二级类别
马克思劳动价值论及其现代形态	F0 经济学	A8 马克思主义等的学习和研究 F0 经济学
学校传染病预防与控制	G4 教育	G4 教育、R1 预防医学卫生学
定性数据统计分析	C8 统计学	C8 统计学、O1 数学
汉字编码的理论与实践	H0 语言学	H0 语言学、TP 计算机自动化技术
诗经与楚辞音乐研究	J6 音乐	J6 音乐、I2 中国文学

可以发现，一些看似预测有误的例子，实际上是对既有图书标引的有益补充。例如根据《中国图书馆分类法第五版（简本）》的设定，C8统计学和O1数学下辖的O212“数理统计”是互见类别[19]。而《定性数据统计分析》一书兼属这两个类别，这样分类不仅是图书管理中两类书籍相互参证的需要，而且有利于提高图书的查全率，促进学科的相互交流。

5　结语

文章着眼于中文图书的细粒度多标签分类工作，考虑到预训练的BERT语言模型的微调（Fine-tuning）特性，提出一种先通过粗粒度分类微调语言模型，在此基础上提取图书表示，再采取惰性学习方法实现细粒度分类的策略。

首先，在面向21大类图书表示学习的单标签分类中，BERT模型在验证集上取得了91.94%的正确率，在遮罩语言模型增量预训练BERT后获得进一步

提升，明显优于前人的LSTM与Fasttext模型。

其次，文章运用带有sigmoid的二元交叉熵损失函数，实现21类图书的粗粒度多标签分类，有92.53%的图书预测出至少1个分类号，有89.98%的图书预测出全部分类号。

最后，文章在微调BERT模型的基础上，采用KNN-ML的方法实现257类的细粒度图书分类。实验表明，经粗粒度分类微调的模型效果明显优于未经微调的预训练模型。在最佳情况下，有75.82%的图书的类别全部预测正确，92.10%图书至少被正确预测了一个类别。从应用角度看，本文提出的分类方法既可以用于图书的自动预分类工作，大大减轻人工标引的负担；也可用于分类号的校对补充，帮助标引者查漏补缺，促进不同学科的沟通与交融。在后续工作中，我们尝试通过加权改进KNN-ML算法，使图书分类系统进一步完善。

参考文献：

[1] 出版商务周报.最新CIP大数据分析，2019图书选题该做什么？ [EB/OL]. (2019-2-28)[2020-04-15].http://www.yidianzixun.com/article/0LOGYM5G.

[2] 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.

[3] 刘高军,陈强强.基于极限学习机和混合特征的中文书目自动分类模型研究[J].北方工业大学学报,2018,30(5):99-104.

[4] 潘辉.基于极限学习机的自动化图书信息分类技术[J].现代电子技术,2019,42(17):183-186.

[5] 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.

[6] 陈志新.分类法研究的十五个问题：我国2009至2016年分类法研究综述[J]. 情报科学,2018,36(6):149-155.

[7] PETERS M E,NEUMANN M,IYYER M, et al.Deep contextualized word representations[J].arXiv e-prints, 2018 :1802-5365.

[8] DEVLIN J, CHANG M, LEE K, et al.BERT: pre-training of deep bidirectional transformers for language understanding[J].arXiv e-prints,2018:1810-4805.

[9] VASWANI A, SHAZEER N, PARMAR N, et al.Attention is all you need[J]. arXiv e-prints, 2017:1706-3762.

[10] SUN C, QIU X, XU Y, et al. How to fine-tune BERT for text classification?[J].arXiv e-prints, 2019:1905-5583.

[11] KINGMA D P, BA J. ADAM: a method for stochastic optimization[J].arXiv e-prints, 2014:1412-6980.

[12] MATTHEW R B, JIEBO L, XIPENG S,et al. Learning multi-label scene classification[J]. Pattern Recognition: The Journal of the Pattern Recognition Society, 2004,37(9):1757-1771.

[13] ZHANG M, ZHOU Z. ML-KNN: a lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007,40(7):2038-2048.

[14] FACEBOOK. Pytorch Docs： Docs > Module code > torch > torch.nn.modules.loss[EB/OL](2019-4-25)[2020-04-15]. https://pytorch.org/docs/stable/_modules/torch/nn/modules/loss.html#BCEWithLogitsLoss.

[15] ELISSEEFF A,WESTON J. A Kernel method for multi-labelled classification[C].In Advances in neural information processing systems,2002:681-687.

[16] ZHANG M L, ZHOU Z H. Multi-label learning by instance differentiation[C]. Proceedings of the 22nd Conference on Artificial Intelligence,2007: 669 - 674.

[17] YANG P, SUN X, LI W, et al. SGM: sequence generation model for Multi-label classification[J]. arXiv e-prints, 2018:1806-4822.

[18] 俞肇元,袁林旺,罗文,等.边界约束的非相交球树实体对象多维统一索引[J].软件学报, 2012,23(10):2746-2759.

[19] 国家图书馆中国图书馆分类法编辑委员会.中国图书馆分类法简本(第五版)[M].北京：国家图书馆出版社,2012:14,107.

蒋彦廷 北京师范大学中文信息处理研究所硕士研究生。北京海淀，100875。

胡韧奋 北京师范大学中文信息处理研究所硕士生导师。北京海淀，100875。

（收稿日期：2019-11-02 编校：左静远，陈安琪）