



开放科学
(资源服务)
标识码
(OSID)

基于 BERT 模型的科技成果中图分类自动标引方法研究

薛钊 刘千祥 吴昌权 李亢 陈永海

中国化工信息中心有限公司 北京 100029

摘要: [目的/意义] 深度学习预训练语言模型 (PLMs) 在科技文献领域分类中的应用效果远超传统自然语言处理技术。科技成果登记数据与科技文献有显著差异, 其简介涵盖项目来源、背景、应用、获奖等多方面内容, 从而大大增加了 PLMs 对科技成果中图分类预测的难度。[方法/过程] 以 BERT 模型 (RoBERTa) 为基础, 在模型集成方法上有所创新, 构建科技成果中图分类自动标引系统。通过引入普适于树形分类体系的解码策略, 将分类问题转化为解码问题, 此举既提升了预测准确率, 又突破了传统分类模型只能在单层级预测的局限, 实现了动态预测。[局限] 受限于所采用的树形分类体系专属解码策略, 暂无法适配不具备树形结构的分类体系。[结果/结论] 通过定制预测链累积概率、终端概率等筛选条件, 该方法可平衡可靠性与分类细致度, 满足不同实际业务需求。

关键词: 科技成果; 自动标引; 深度学习; BERT 模型; 解码策略

中图分类号: G254.11; TP18; G35

Research on Automatic Indexing Method of Chinese Library Classification of Scientific and Technological Achievements Based on BERT Model

XUE Zhao LIU Qianxiang WU Changquan LI Kang CHEN Yonghai

China National Chemical Information Center Co., Ltd., Beijing 100029, China

Abstract: [Objective/Significance] Application of deep learning pre-trained language models (PLMs) in the classification of scientific and technological literature outperforms traditional Natural Language Processing techniques. There are significant differences between the scientific and technological achievements and scientific and technological literature. The introductions of the former cover various aspects such as project sources, backgrounds, applications, and award-winning information, which

作者简介 薛钊 (1983-), 通信作者, 博士, 工程师, 主要研究方向为深度学习、NLP, E-mail: xuez@cncic.cn; 刘千祥 (1972-), 硕士, 高级工程师, 主要研究方向为数据资源建设、知识与情报工程等; 吴昌权 (1964-), 硕士, 教授级高级工程师, 主要研究方向为数据资源建设、知识与情报工程等; 李亢 (1986-), 硕士, 工程师, 主要研究方向为数据挖掘与数据资源建设; 陈永海 (1974-), 硕士, 工程师, 主要研究方向为数据标引。

引用格式 薛钊, 刘千祥, 吴昌权, 等. 基于 BERT 模型的科技成果中图分类自动标引方法研究 [J]. 情报工程, 2025, 11(5): 3-12.

significantly increases the difficulty of predicting the chinese library classification using PLMs. [Methods/Processes] This work is based on the BERT model (RoBERTa) and innovates in the model integration method to construct an automatic indexing system for the chinese library classification of scientific and technological achievements. By introducing a decoding strategy, which can be generally applied to the tree-structured classification system, the classification problem is transformed into a decoding problem. This not only improves the prediction accuracy, but also enables dynamic predictions to the required levels. [Limitations] Constrained by the adopted decoding strategy exclusive to tree-structured classification system, this method cannot be directly adapted to classification system without a tree structure. [Results/Conclusions] By customizing predict conditions such as the cumulative probability of the prediction chain and the terminal probability, this method can balance the trade-off between reliability and classification fineness to meet practical needs.

Keywords: Scientific and Technological Achievement; Automatic Indexing; Deep Learning; BERT Model; Decoding Method

引言

科技成果登记统计是全国科技统计工作的重要组成部分,是掌握国家科技发展状况的重要途径。积累的科技成果登记数据,对科技宏观管理及科技成果转化意义重大,为科技成果的推广应用及产业化服务提供了有力的数据支撑。科技成果数据覆盖全行业领域,在数据加工和实际应用中,通常借助中图分类法对科技成果进行分类标引,以实现数据的快速筛选定位。

《中国图书馆分类法》(简称《中图法》)是国内广泛采用的大型综合性分类法。最新版第五版涵盖约6万个类目,每个类目均对应唯一的中图分类代码。《中图法》采用树状层级结构,以空节点为根,首层级设22个学科大类。类目数量在第五、六层达到峰值,各含约1.6万类目。学科分类内涵呈父子节点关系,如TP31(计算机软件)包含子节点TP319(专用应用软件)。此外,《中图法》部分类目学科内涵相近^[1],易造成分类混淆,但也赋予分类过程一定“弹性”。

中图分类层级多、类目数量庞大,对大批量、全领域的科技成果数据标引而言,是一项艰巨

挑战。中图分类标引是科技成果数据加工的关键环节,其标引质量对科技成果的领域标识至关重要。传统的科技成果数据中图分类标引采用人工集中标引方法,效率较低。而且,因集中标引通常由多人执行,在分类的细致程度和相似类目认定上,难以统一标准。随着计算机技术的发展,降本、提质、增效的业务需求推动人们不断探索自动化标引的技术手段^[2-9]。

1 研究现状

早期的尝试主要使用依赖关键词及词串的相关算法,如语料库支持下的支持度分析^[2]和机器学习算法^[3-6]。传统机器学习算法(如朴素贝叶斯、支持向量机等)大多需要文本分词及特征抽取,导致在信息输入模型之前就引入了不确定性,同时不得不设计复杂的特征工程和系统架构^[4]。自然语言中字词以序列呈现,在相互关联之中蕴含大量信息,即通常所谓的语义。相较于依赖词汇层面的算法,将文本作为序列输入至深度学习模型中,以捕捉其内在的语义信息,这在方法论上是显著的进步。这一时期较多采用的深度学习模型包括LSTM^[10]、

TextCNN^[11]等,将注意力机制引入模型架构通常会进一步提升效果。相比传统机器学习算法,系统架构大幅简化,预测效果也不断提升^[7-9]。2018年,谷歌公司提出并开源了BERT模型^[12],提出了一种以少样本学习的方式解决文本序列分类问题的方法。BERT模型基于Transformer架构^[13],拥有全局注意力机制,具有更强的语义理解能力。同时,开源的BERT模型已经利用大量数据做了预训练,从大数据中学习到了基础语义,下游用户通常只需在自己掌握的少量数据上微调就可以取得很好的效果。

罗鹏程等^[14]将BERT模型应用于中文情报和文献分类,以21个人文社科领域为分类目标进行实验,结果表明基于预训练语言模型(PLM)的算法超越了传统机器学习和深度学习所能达到的效果。沈立力等^[15]提到,基于BERT构建的分类模型在中文期刊上执行分类任务,在《中图法》的特定大类和层级上取得了不错的效果。同时明确指出,“面对大数量级、多类目分类问题时,目前还未从文献调研中发现BERT模型在类似《中图法》这样极大数量类目分类问题上有实际可投入生产的成果”。选择中图分类的子集进行实验,本质上减少了预测结果的取值空间,降低了预测难度。将中图分类标引简单地视为分类问题,则只能在特定层级微调模型并执行预测,这进一步限制了算法的应用场景。如何构建全领域的、动态层级预测的中图分类自动标引系统,仍需在方法上有所突破。

BERT模型用于下游分类任务有多种使用方式,可以直接将[CLS]位置上的输出作为分类层的输入,也可以将所有位置的输出整合起来作为分类层的输入。李湘东^[16]使用了另外一

种方式,将BERT每一层的输出整合起来作为分类层的输入,直接利用不同深度的表示进行分类预测,并声称取得了较好的效果。

2 研究方法

本文旨在通过设计适用于科技成果数据的中图分类自动标引算法,最终构建一个能够在实际生产中应用的中图分类自动标引系统。因此,需打破常规仅在某些固定的层级上给出预测的局限^[15],采用模拟人工标引时动态选择分类层级的方式。同时对模型预测的准确率提出了较高的要求,最大程度减少人工干预及检查的需求,最终达到“以假乱真”的程度。

BERT模型在中文文献分类任务中已有出色表现,并且科技成果数据具有文本特征,因此适用于基于BERT模型衍生出来的方法。与科技文献相似,科技成果数据的标题、简介、关键词中均蕴含其学科领域信息。科技成果高度凝练的标题中包含的信息密度较大,较为详细的长文本简介中包含的信息总量较多。同时,科技成果数据又有其独特之处,比如简介中通常包含项目来源、项目背景、应用情况、获奖情况等,是对科技成果多方面的介绍,而科技文献的简介会高度聚焦于研究内容。据此推测,即便采用相同的模型和算法,对于科技成果学科领域的预测效果也会打折扣。鉴于科技成果数据的特点,本研究设计了具有针对性的数据处理方式生成模型训练数据集。同时在预测阶段,将单项科技成果转化为多个输入,对预测结果的集成可以增加预测结果的可靠性,基本流程如图1所示。

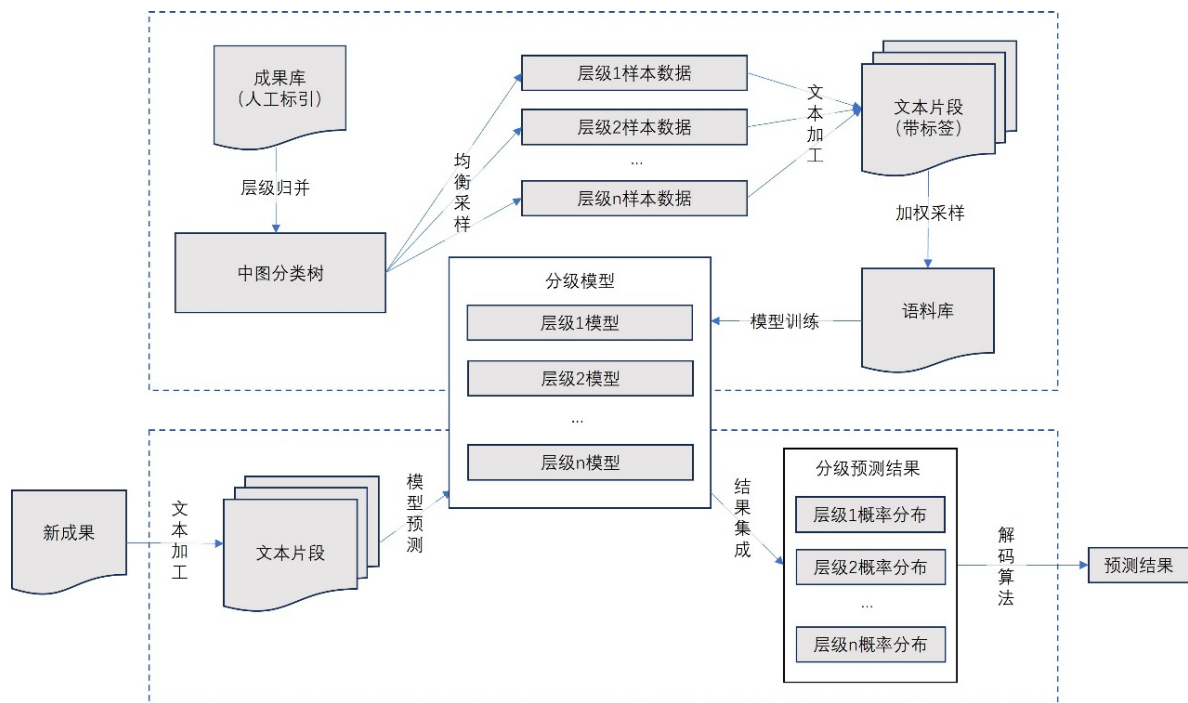


图1 系统基本流程

为了实现动态预测，即最终预测结果可能处于中图分类的不同层级，设计采用了分层预测+预测链判定的方式。首先，逐级训练独立的分类模型，执行预测时，模型将在每个中图分类层级上输出独立预测概率。然后，设计一种类似生成式大语言模型（LLM）的解码策略：引入温度参数调整预测概率的相对大小；引入参数 top-n 和 top-p 对分量进行筛选；考查筛选后的结果中由父子关系组成的“链”，将它们类比为 LLM 生成的文本。如此便将单纯的文本分类问题转化为了序列解码问题。预测链对应着中图分类树上从根节点开始的路径。能够形成预测链，说明各层级独立训练的模型达成了“共识”，于是预测链的累积概率就可以很好地反映预测的可靠性。为了利用集成学习带来的提升效果，尽管可以在同一个底层 BERT 模型上构建能够在多个层级执行预测的模型，即

让模型具有多个输出头，但各层级相互独立的模型更能发挥集成学习的优势。为了进一步保证可靠性，设置了截断阈值，预测链的终端则为潜在的最终预测结果。为了限制链的长度影响过大，即模型倾向于输出最长预测链，还引入了长度惩罚因子，其值由实验测试确定。在解码策略的支持下，系统不仅实现了动态预测的目的，而且由于集成学习的影响，相比单一模型还极大地提高了预测的可靠性。

相较于最初的 BERT 模型，RoBERTa 模型对训练策略和模型细节做了少量改进^[17]。综合考虑模型效果、训练及推理效率，选用了 RoBERTa 的 base 版本作为基础模型。但本研究并不将通过精益求精的训练提高单个基础模型的能力当作研究重点，而是将注意力集中在模型集成及解码算法的设计。项目工程中还预留了集成其他 BERT 模型的接口，不同开源 BERT

模型通常是由不同的数据集及流程训练而来，学习的基础语义彼此间存在差异，据此推测，集成不同模型可以再次提高系统的准确率。

3 设计与实现

针对科技成果数据的特点，通过特征工程从历史数据获得模型训练数据集。模型主体采用常规的 BERT 分类模型，重点放在多模型分层级预测后解码策略的设计上。

3.1 特征工程

选取国家科技成果库中 2011 至 2022 年共 12 年的历史数据用于构建模型训练数据集，精选出 65.9 万项科技成果，然后再从中划分出独立的训练集、验证集和测试集。数据集中与学科领域相关的字段包括标题、简介、关键词、中图分类代码。

标题和关键词都是短文本，标题大多是较为完整的句子，关键词则为短词，均对学科分类有较强的预测能力。简介是长文本，与学科领域直接相关的研究内容介绍通常出现在前部，而转化应用、获奖情况等不与学科领域有强相关性的内容通常出现在后部。通常 BERT 模型支持的最大序列长度有限^[12]，简介长度大多超出限制，因此需要对简介进行切分。针对简介的特点，对切分后的片段设计了加权采样，如图 1 所示，将标题与从简介片段中采样得到的文本片段拼接起来，形成模型的单条输入数据。输入特征没有引入关键词，一是因为关键词不构成句子，对模型语义理解未必有利，二是考虑到关键词大多已经出现在简介当中，无需重复计入。

将数据集中标注的中图分类代码依照中图分类树向上归并，得到每项科技成果的各层标签。在中图分类的各个层级上，数据集的类目数量分布是不均衡的，甚至某些类目没有任何数据。例如，科技成果多属于应用技术类，属于 T 分类的成果数量要远远多于其他一级分类的成果数量。为缓解这种不平衡带来的影响，对中图分类树进行了修改，将 T 分类下属各类目均上提一个层级。除此之外，辅助设计了采样算法，进一步地降低了数据不均衡性。

每项科技成果至少标有 1 个中图分类代码，有大约 20% 的科技成果标有 2 个中图分类代码。科技成果可能属于 2 个或多个类目，它们通常为跨领域的成果。实际上，到分类较为细致的时候，科技成果从不同的视角理解，大都可以标注多个中图分类。这种“弹性”将带来两方面的影响：一是增加了原始数据人工标注的难度，人们在区分度较低的选项中做出选择时更容易出错；二是为模型预测准确率提升提供了缓冲。从后续的解码过程看，跨领域科技成果对形成充足数量的预测链至关重要。

中图分类树中节点路径终止时的深度各不相同，如 S17（农业地理学）在 3 级分类终止，其下没有任何子类目。如果将原本属于 S17 的科技成果输入到后续层级的模型中进行预测，因其训练时从未见过此类数据，将导致模型给出不可靠的预测。这种不可靠性会严重影响基于预测链的解码过程。针对此情况再次修改了中图分类树，每个层级添加 OTHER 类目，将当前层级之上已经终止的类目所对应的数据收集起来，作为当前层级 OTHER 类目的训练数据，避免模型给出无效的强制分类。

3.2 模型架构

系统采用了最常用的模型架构，即将 BERT 模型第一个输出作为线性分类层的输入，调整为目标维度后输入 Softmax 层得到各分量的预测概率。模型在线性层之前添加了 Dropout 层，

以防止过拟合。如前所述，在中图分类树每个层级训练独立的模型。在中图分类树上做统计，发现科技成果集中分布在第 4~6 级，其中第 4 级已超过 43%，第 7 级及以下累计约 2%。统计结果表明只需在中图分类的 1 至 6 级训练独立的分类模型，系统模型架构如图 2 所示。

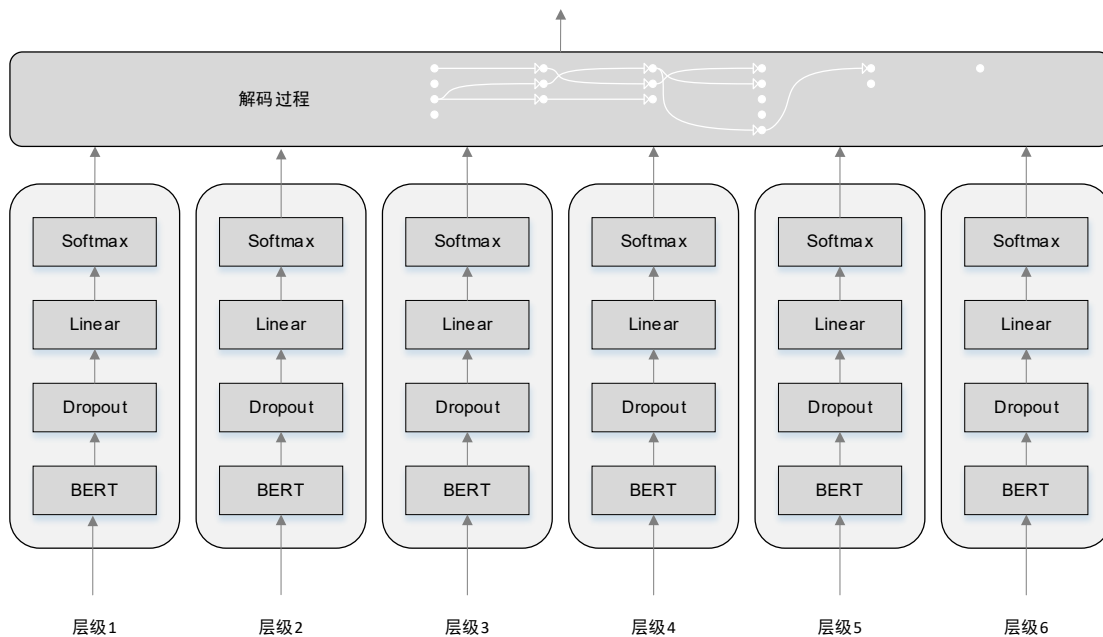


图 2 系统模型架构

模型微调的学习率参数为 $3e-5$ ，选用 Adam 损失函数，各层级模型在 8 个周期左右达到最佳。

3.3 解码策略

在执行预测时，将单项科技成果的简介切分之后与标题拼接，形成数个文本序列。将模型的输出向量集成起来有多种方式，如取平均值、取最大值以及投票。实验测试发现，取平均值的方式效果最好。

根据训练用数据集的统计结果，在预测标签中剔除了从未出现的中图分类代码。即便如此，各个层级的类目数量仍然很多，比如第 5

级有约 5500 个类目。如果考虑所有的分量，会由于计算量过大导致效率较低。此时采用了类似生成式大语言模型的解码策略，引入了温度参数、top-n、top-p 来控制生成的多样性和降低计算的复杂度。预测概率向量的相对大小受温度参数的影响，top-n 和 top-p 则对需要考虑的分量做了筛选。

每个层级保留 1 至数个预测结果，从中可以收集所有的预测链。所谓预测链即满足父子关系节点组成的序列，对应着中图分类树上的一条连续路径。以科技成果“新型高速模块全自动贴标技术研发”为例，各级模型输出中图分类代码及其对应的概率如图 3 所示。

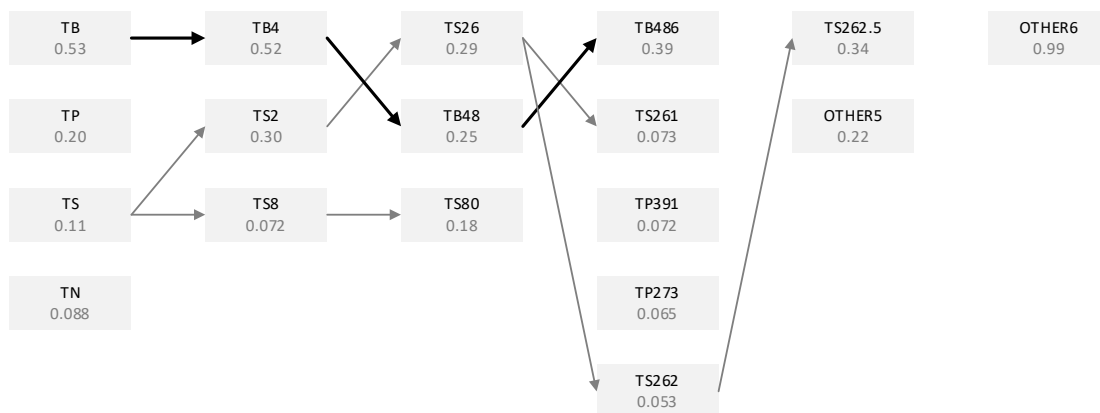


图 3 中图分类预测链

预测链的累积概率为：

$$P_{\text{chain}} = \prod_i P_i(\hat{y}_i) \quad (1)$$

各层级的模型是独立训练的，因此累积概率能够比模型的单个预测更加稳定可靠。考虑累积概率的对数值，并引入长度惩罚因子：

$$\log P_{\text{chain}} = \frac{1}{L_{\text{chain}}^\alpha} \sum_i \log P_i(\hat{y}_i) \quad (2)$$

其中 L_{chain} 是序列长度， α 为长度惩罚因子。

惩罚因子影响预测链的平均长度，最终反映在预测结果的细致程度上。

图 3 中共存在 4 条预测链，由公式 (1) 分别计算它们的累积概率，长度惩罚因子根据后续测试及优化结果选用经验值 $\alpha=2.7$ ，如表 1 所示。

表 1 预测链实例分析

预测链	1 级	2 级	3 级	4 级	5 级	对数概率
1	TB 一般工业技术	TB4 工业通用	TB48 包装工程	TB486 包装机械设备		-0.0857
2	TS 轻工业、手工业	TS2 食品工业	TS26 酿造工业	TS261 酿酒工业		-0.172
3	TS 轻工业、手工业	TS2 食品工业	TS26 酿造工业	TS262 各种酒及其制造	TS262.5 啤酒	-0.112
4	TS 轻工业、手工业	TS8 印刷工业	TS80 一般性问题			-0.337

根据预测链累积概率对数值，最佳预测结果为第一行预测链的终端 TB486（包装机械设备）。从这项成果的简介中可以看出，该设备主要用于啤酒工业的包装过程，从而可以解释其他几条预测链的存在。

4 实验结果分析

首先在历史资源数据划分出的测试集上对中图分类自动标引系统进行了测试。考虑到测试集的标引质量以及科技成果中图分类标引的弹性，还对系统标引结果进行了人工审核。针

对模型测试和人工审核过程中遇到的典型问题，对算法进行了迭代优化。

4.1 模型测试

模型测试设置了4个测试指标：（1）每个层级概率最大的预测是否与人工标引匹配；（2）每个层级概率前3的预测是否包含人工标引中的至少一项；（3）引入解码策略后的预测结果是否与人工标引匹配；（4）解码策略选择的预

测链与人工标引形成的预测链在中图分类树上是否重合或覆盖。测试指标（4）对于测试数据均标有1或2项中图分类代码情况，与预测有一项相同则视为匹配。

用于测试的测试集包括12912项科技成果历史数据。人工标引的层级有深有浅，考虑到中图分类可以向上归并，测试样本的数量随着层级深度增加逐渐减少。模型测试结果如图4所示。

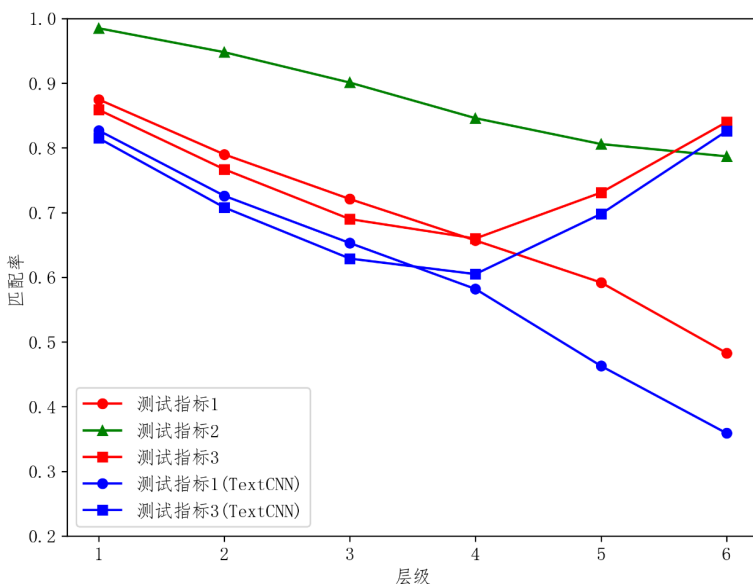


图4 模型测试结果

图4中红色和绿色线条为前3项测试指标的测试结果。可以看出，随着层级加深，最大概率匹配（测试指标1）的比例持续下降。这是由于类目数量随层级加深而增大，且学科内涵的区分度越来越小，做出正确预测就变得更加困难。如果考虑预测前3概率匹配（测试指标2）情况，其比例相较于最大概率匹配有大幅提升，这一特性为通过设计解码过程提高预测准确率提供了基础。引入解码策略后匹配率（测试指标3）在低层级1~3级并未有比最大概率匹配更好的表现，但是在高层级4~6级有

大幅提升，扭转了持续下降的趋势。匹配率的提升反映的是模型集成带来的影响，有更多的模型达成了共识，预测结果变得更加可靠。在中图分类的高层级拥有更好的表现，正是科技成果标引工作所需要的特性。预测链匹配（测试指标4）反映了人工标引与自动标引在除分类细致程度之外达成完全一致的情况，测试的数值为0.601。

图4中蓝色线条为同等条件下在科技成果数据集上构建的TextCNN模型的测试结果。TextCNN为浅层神经网络架构，且未经过预训

练，测试结果表明其所能达到的匹配率略低于基于 BERT 构建的模型。使用解码策略之后，在高层级同样带来了预测匹配率的提升，这反映了该方法具有普适性。

4.2 人工审核

人工审核由专业的标引人员按照统一的数据质量标准进行，本次人工审核设置了 2 个测试指标：可用和不可用。同时，通过对预测链终端节点的预测概率设定阈值，系统也可以给出可用和不可用推断，在实际标引工作中可以据此剔除少量不可靠的预测数据，直接进行人工处理。

表 2 人工审核结果

系统	人工	
	可用	不可用
可用	4812	19
不可用	35	10

用于人工审核的数据集为从最新登记的科技成果中随机提取的 4876 项，它们未做任何加工处理。测试结果如表 2 所示，据此可以计算出准确率（accuracy）为 0.989。系统标注为可用的预测中，可用率达到了较高水平的 0.996。人工审核的可用率大幅超过模型在测试集上测试时的匹配率（图 4 测试指标 3），其中的差异可由测试集的标引质量以及科技成果中图分类标引的弹性来解释。

5 结语

随着 PLMs 的发展，人们能够在有限的条件下，比如仅掌握少量数据，也可以在 NLP 任

务中取得较为理想的效果。本文将 PLMs 应用于科技成果的中图分类标引任务中，基于 BERT 模型（RoBERTa）构建了用于生产环境的中图分类自动标引系统。针对科技成果数据的特点，制定了从原始科技成果数据集获得模型训练数据集的数据处理及采样方法，对模型使用的中图分类树也做了必要的调整。引入解码算法是自动标引系统能够达到理想效果的关键，此举将单纯的分类问题转化为解码问题。解码算法解决了以往分类模型只能局限于单一级别执行预测的问题，能够实现动态预测，满足实际业务需求。同时，通过为每个中图分类层级训练独立的预测模型，在一定程度上发挥了集成学习的力量，借助预测链判定形式提高最终预测的准确率。这些方法共同保证了科技成果中图分类自动标引系统的预测效果，使其在科技成果数据资源治理中发挥了重要作用。

由于中图分类类目众多、数量庞大，针对其进行的自动预测被视为一项较为困难的任务。鉴于科技成果数据所具有的特点，以及原始数据集人工标引的质量状况，致使针对科技成果的中图分类自动标引任务更为困难，在中图分类单个层级上训练得到的模型其表现无法与在期刊文献上的同类任务相媲美^[15]。本研究的测试和实用效果表明，引入解码策略能够极大地提高模型的最终预测能力，不仅可在针对期刊文献等中图分类标引任务中推广应用，而且可以尝试将该方法推广至任何树形分类体系的文本分类任务，如专利 IPC 分类体系。

参考文献

- [1] 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.
- [2] 侯汉清, 薛春香. 用于中文信息自动分类的《中图法》知识库的构建[J]. 中国图书馆学报, 2005(5): 82-86.
- [3] 薛春香, 夏祖奇, 侯汉清. 基于语料和基于标引经验的自动分类模式比较[J]. 南京农业大学学报(社会科学版), 2005(4): 85-92.
- [4] 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.
- [5] 杨敏, 谷俊. 基于 SVM 的中文书目自动分类及应用研究[J]. 图书情报工作, 2012, 56(9): 114-119.
- [6] 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.
- [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] 孔洁. 基于深度学习与《中国图书馆分类法》的文献自动分类系统研究[J]. 新世纪图书馆, 2021(5): 51-56.
- [10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [11] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
- [12] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017(30): 6000-6010.
- [14] 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.
- [15] 沈立力, 姜鹏, 王静. 基于 BERT 模型的中文期刊文献自动分类实践研究[J]. 图书馆杂志, 2022(5): 109-118.
- [16] 李湘东. 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.
- [17] CUI Y M, CHE W X, LIU T, et al. Pre-Training with whole word masking for Chinese BERT[J]. arXiv preprint arXiv: 1906.08101, 2019.