基于 BERT-RL 的红色文献元数据标注研究*

□武帅　何琳　杨秀璋　孙绍丹　桂思思

摘要　红色文献普遍存在数量庞大、分散且类型复杂多样等特点，使得基于零基础训练的大语言模型较难实现高效应用。如何利用少量已标注数据来微调大型语言模型，从而实现红色文献领域元数据的自动化标注，是当前亟须解决的关键问题之一。文章设计了一种基于增量学习的大语言模型微调方法，并结合深度学习技术，实现了红色文献元数据表示体系的自动化标注。首先，根据红色文献的内容特征，将其归纳为事件叙述、人物评价、情感表达、文献摘录和场景描述五类文本数据集；其次，采用增量学习方式对大语言模型进行微调；最后，构建融合 BERT-RL 预训练模型和 CNN-BiGRU-MHA-CRF 的自动化标注模型，以实现对红色文献元数据的自动化标注。研究结果表明，目标领域数据集的数据特征多样性有助于提升基于增量学习微调大语言模型的效果。

关键词　增量学习　红色文献　元数据表示体系　自动化标注　BERT-RL
分类号　G251.2
DOI　10.16603/j.issn1002－1027.2025.04.009
引用本文格式　武帅，何琳，杨秀璋，等. 基于 BERT-RL 的红色文献元数据标注研究[J]. 大学图书馆学报，2025，43(4)：76－89.

1　引言

红色文献作为红色资源的重要组成部分，系统记载着中国共产党带领中国人民顽强奋斗、革命解放的历程[1]。构建大规模红色文献知识库，实现文献中知识实体的细粒度标注及其关联关系的深度挖掘，是推进红色文献研究数字化与智能化转型的重要基础[2]。然而，红色文献具有类型多元、结构异构和语义关联复杂等特征[3]，为红色文献的大规模自动化处理带来一定影响。当前，以大语言模型为代表的人工智能技术得到快速发展，带动着人文研究从传统人文范式向数字人文研究范式的转型[2]，这为红色文献知识库建设提供了新的技术支撑。因此，针对红色文献的特点，综合利用大语言模型技术提升红色文献基础数据智能化处理效果，是全面推动红色文献知识获取、挖掘及生成质量的关键[4]。

红色文献普遍存在数量庞大、分散且类型复杂多样等特点，使得基于零基础训练的大语言模型较难实现高效应用。微调技术作为迁移学习的技术核心，能够更好地将模型预训练期间获取的大量知识应用于特定领域，是处理具有特定语境或专业知识场景的关键[5]。基于此，如何在有限资源的前提下，通过微调技术提升传统预训练模型在红色文献领域的适配性，确保红色文献领域专业化知识获取的质量，是当前亟须解决的关键问题之一。

针对上述问题，本研究初步尝试以增量学习的方式微调出适用于红色文献领域的预训练模型(Bidirectional Encoder Representations from Transformers-Red Literature，BERT-RL)，为实现红色文献知识库规模化建设提供核心技术支持。

2　研究现状

2.1　红色文献标注的研究现状

目前，红色文献标注的研究多集中于理论层面的体系架构探讨，实践层面的应用研究相对较少。在理论层面的体系架构方面，主要探讨在复用已有

* 国家社会科学基金重点项目“雨花英烈文献知识组织及智能内容生成研究”(编号：23ATQ012)的研究成果之一。
通讯作者：何琳，OCRID：0000-0002-4207-3588，邮箱：helin@njau.edu.cn。

本体模型的基础上，设计针对红色文献特征的领域本体模型。例如，基于 BIBFRAME 2.0 的书目本体和上海记忆本体模型，对上海图书馆的红色文献进行知识重组[6]。此外，基于 BIBFRAME、CIDOC-CRM、SEM 和 FOAF 等本体框架，还设计了皖西红色文献知识本体[7]、北京香山红色档案本体[8]、中共七大文献本体[9]以及新四军苏浙军区红色文献本体等[10]。

在实践层面，学者们尝试运用深度学习等技术抽取红色文献中蕴含的相关实体及实体关系。例如，运用规则抽取的方式实现对雨花英烈各类型文献的人物属性抽取[11]，运用 BERT-BiLSTM-CRF 模型实现对陕甘宁边区政府公文实体的自动化识别[12]以及《全国重要革命遗址通览》文本中实体的自动化抽取[13]。在实体关系抽取中，也探究了通用文本处理工具的适用性，例如，使用哈尔滨工业大学云平台对大别山革命文物资源进行抽取[14]。

通过对现有研究的系统梳理发现：在理论层面，现有红色文献标注体系主要聚焦于单一类型或特定文本的语义关系揭示，所设计的标注框架虽然能够较好地适应特定文献类型的标注需求，但在扩展至大规模、多类型的红色文献自动化标注任务时，普遍存在适应性和可扩展性不足的问题。在实践层面，当前研究大多局限于将通用模型直接应用于某一类红色文献的自动化标注，该类研究虽验证了通用模型在红色文献处理中的基础适用性，但未能深入结合红色文献特有的内容特征和知识结构，缺乏对大语言模型进行领域适配性优化的系统性探索，从而限制了模型在红色文献处理任务的准确性。

2.2 大语言模型的微调研究

在现实应用中，知识标注的场景多以特定语境或专业知识为主[15]，通用大语言模型[16]往往无法准确捕获关键语义信息，存在召回率较低的现象。微调作为连接大语言模型和特定应用场景的桥梁[17]，通过特定数据集对大语言模型进行领域性适配性优化，提升模型在特定语境或专业知识场景下的处理能力，有效增强模型识别效果[4]。微调方法通常可分为：监督微调和无监督微调两种。[18]。其中，监督微调通过在特定任务集上对初始参数进行调整，提升原模型的准确性，被广泛运用于自然语言处理的具体任务中[19]。无监督微调受限于语料不包含标记，需模型自行理解数据，仅适用于大语言模型的二次预训练环节[20]。

在大语言模型中，GPT 模型[21]作为首个在大量文本上使用无监督微调的大语言模型，虽实现了模型微调，但性能提升并不显著。GPT-2 模型[22]在其基础上对参数和数据集的规模进行扩充，并通过监督微调的方式在多项任务中取得显著提升效果。BLOOM 系列大语言模型[23]尝试在 130 亿字符文本集上微调出最优模型。ChatGPT 模型[24]在 GPT-3 模型的基础上通过人类反馈的三阶段强化学习算法（Reinforcement Learning from Human Feedback，RLHF）的监督微调，使其以符合人类偏好的方式输出结果。相较于上述参数量大的模型，BERT 模型[25]以较小的参数规模，受到众多学者的青睐，先后推出 RoBERTa[26]、ALBERT[27]等小型的大语言模型。基于上述模型，开发了 BioBERT[28]、SciBERT[29]、GuwenBERT[30]等模型，实现在生物医学、科技文献和古籍文本等场景下低成本的全参数微调。

通过文献梳理发现，大语言模型的微调多侧重于大规模参数语言模型的迭代优化和小规模参数语言模型的领域微调，而从数据层面探讨领域微调可行性的研究甚少，缺乏考虑如何微调出某一特定语境或专业知识场景下模型效果的理论知识。

2.3 研究现状评述

现阶段基于大语言模型的微调研究仅证明微调后模型的效果优于原模型，并未考虑特定语境或专业知识场景下标注数据规模的局限性，未能充分考虑数据规模与大语言模型微调效果间的关联性，忽略了样本数据规模对大语言模型微调准确性的影响。因此，本研究认为如何确认特定语境或专业知识场景下所需的样本数据规模，是实现大语言模型高质量微调的前提条件。

基于此，本研究尝试以增量学习的方式微调出适用于红色文献的大语言模型，并构建具有迁移学习能力的标注模型，以实现红色文献的自动化标注。具体包括：(1)相较于传统以文献为单元的外部特征著录作为数据集划分特征，本研究以红色文献的内容特征划分数据集，探究大语言模型微调所需的样本数据类型。(2)针对红色文献的内容特征，设计自动化标注模型，进一步提升红色文献元数据标注的精度。

3 红色文献元数据标注研究框架

3.1 红色文献元数据标注的框架

本研究将课题组前期研究的 MRS 6 元数据体系[31]所标注的红色文献作为 BERT-RL 的微调数据集，结合增量学习的思想设计出红色文献元数据标注的研究框架（见图 1）。

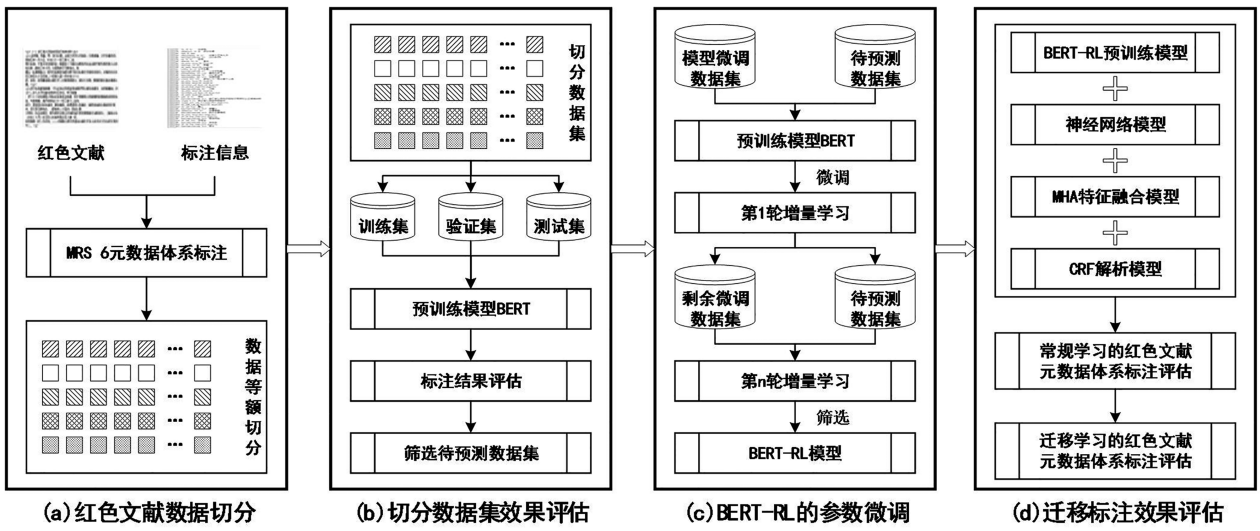

图 1 基于 BERT-RL 的红色文献元数据标注的研究框架

（1）红色文献数据切分。将课题组前期研究的 MRS 6 元数据体系所标注的红色文献按照文献内容特征划分为五类文本数据集，并将每类文本数据集等额切分为十组微调数据集。

（2）切分数据集效果评估。将切分的各类型微调数据集分别按照 8∶1∶1 的比例分为训练集、验证集和测试集，输入 BERT 预训练模型进行效果评估，筛选出实验效果最低的微调数据集，作为评估 BERT-RL 模型微调效果的数据集。

（3）BERT-RL 模型的参数微调。对 BERT 预训练模型进行多轮增量学习，并将每轮微调效果最佳的参数组合继承至下一轮增量学习的预训练模型，迭代训练至微调完全部数据，筛选出最佳效果的参数组合继承至 BERT-RL 模型。

（4）迁移标注效果评估。构建了融合预训练 BERT-RL、卷积神经网络（CNN）、双向门控单元（BiGRU）、多头自注意力机制（MHA）和条件随机场（CRF）模型的 BERT-RL-CNN-BiGRU-MHA-CRF（简称 BERT-RL-CBMC）红色文献自动化标注模型，分别对已标注（常规学习）和未标注（迁移学习）的红色文献进行 MRS 6 元数据体系标注评估。

3.2 增量学习

以增量形式提供红色文献数据，使 BERT 预训练模型能够快速捕获红色文献的领域特征知识，提升模型微调效果。通过多轮迭代式增量学习，微调出适用于红色文献领域的 BERT-RL 预训练模型，其工作原理详见图 2。

3.3 红色文献元数据标注模型

结合红色文献标注的需求，本研究构建了 BERT-RL-CBMC 模型，以识别红色文献的元数据，模型网络结构详见图 3。

（1）BERT-RL 预训练层。经过增量学习的 BERT-RL 模型不仅保留了 BERT 模型的双向 Transformer 架构，实现对红色文献的字向量、词向量和句向量的三层语义特征嵌入；[Mask]机制补全缺失语义；还通过增量学习的微调方式优化模型处理红色文献的能力。

（2）神经网络层。本研究创新性构建融合 CNN 和 BiGRU 的神经网络层提升模型整体的识别效果。CNN 网络主要由卷积层和池化层构成，卷积层通过卷积核滑动搜集红色文献短句信息来捕获多个连续特征，池化层通过向量采样降低模型的计算量，最终识别出红色文献的局部关键特征。BiGRU 网络捕获红色文献的正向和反向信息，更全面地挖掘红色文献的语义特征，尤其是文本时序的全局语义特征和双向依赖关系。GRU 作为循环神经网络的变种，其网络结构虽与 LSTM 类似，但比 LSTM 模型的参数量更少，有效提升红色文献元数据标注模型的训练效率。

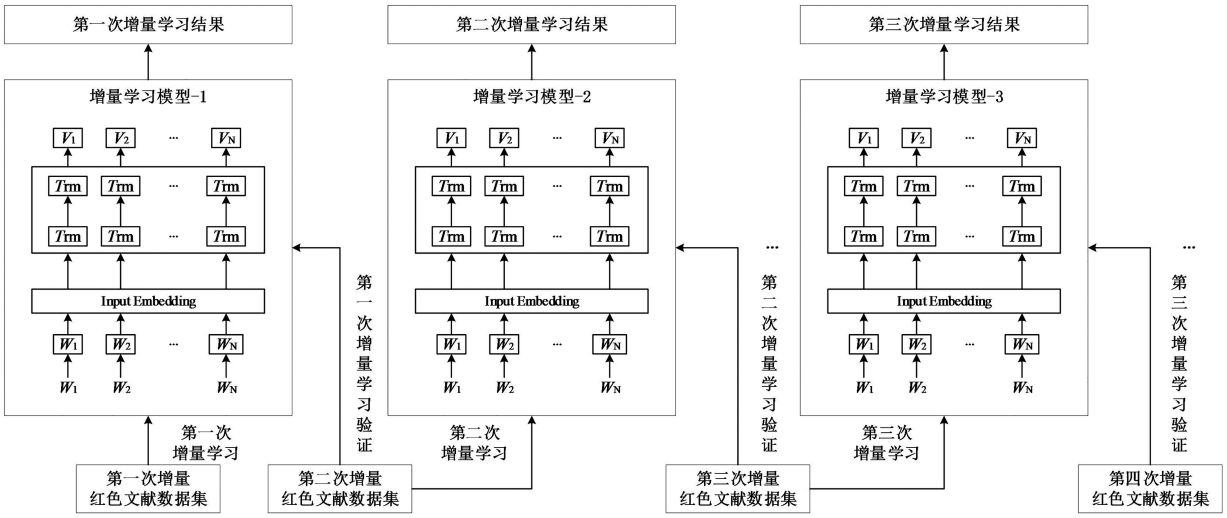

图 2　基于增量学习的预训练模型 BERT 微调流程概念图

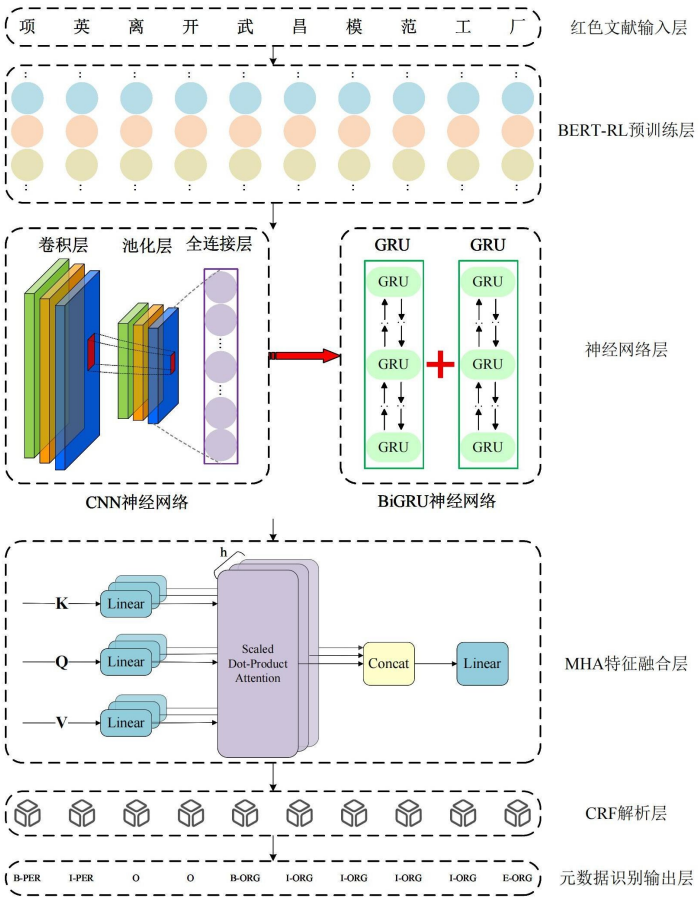

图 3　红色文献元数据标注模型的网络结构图

（3）MHA 特征融合层。考虑到传统向量叠加的融合方式仅能获取不同模型获取的语义特征信息，并不能实现语义特征的融合，本研究利用 MHA 模型的多头注意力并行能力将 BERT-RL 预训练模型和 BiGRU、CNN 网络提取的特征进行融合，有效突出红色文献文本序列的语义关联信息和时序特征，弥补循环神经网络在建模长期依赖信息时易发生梯度爆炸（消失）的问题。

(4)CRF 解析层。由于上述模型均无法感知实体及相邻标签间的依赖关系，本研究利用 CRF 模型计算标签间的转移关系以及标签序列概率，求解出概率最大路径以确定红色文献不同汉字的节点标签，实现对红色文献的元数据标注。

至此，BERT-RL 预训练模型通过增量学习的方式增强 BERT 预训练模型对红色文献的处理能力，提升对待标注元数据词边界的界定准确性；CNN 网络通过卷积核滑动获取元数据的关键特征；BiGRU 网络通过正反向门控循环单元获取元数据的位置特征；MHA 模型融合上述特征，进一步强化元数据的关键特征；CRF 模型通过最大似然估计计算出最佳序列标注结构，以实现对红色文献高精度、细粒度、深层次的元数据标注。

4 实验结果与分析

为验证 BERT-RL 模型与 CNN-BiGRU-MHA-CRF 模型(简称 CBMC)的领域适应性和标注性能，本研究设计了一套系统的多维度评估方案，包括：基于增量学习的预训练模型性能评估、常规学习的元数据体系标注评估、迁移学习的元数据体系标注评估以及基于增量学习的预训练模型消融实验。

在实验环境配置方面，本研究采用 Keras 2.6.0 框架实现深度学习模型架构，基于 TensorFlow 2.6.0 平台构建预训练学习模型，编程语言选用 Python 3.9 版本。硬件配置采用 Intel Core i9-13700K CPU、NVIDIA GeForce RTX 4090 GPU 和 64G 内存。为确保实验结果的可靠性，所有对比测试均在统一的软硬件环境下完成，并设置了多组基线模型进行对比分析。

模型参数设定方面，词向量维度设为 500，卷积神经网络的卷积核设为 128，BiGRU 和 BiLSTM 的前向和后向隐藏层数量均设为 128，初始学习率设为 0.001，注意力机制头数设置为 100，优化器均选择 Adam 求解最小损失，训练批次设为 20，最大迭代次数设为 100，批量梯度下降数设为 128，模型均设置相同的激活函数(ReLU)与归一化方法。

为确保对比模型实验结果的公平性，避免单次错误实验结果造成的片面性判断，所有模型均采用十次交叉实验结果的算术平均值作为模型性能评估值，以提升对比实验结果的真实性与普遍性。

4.1 实验数据集

(1)标注体系

现阶段，学界和业界虽已完成大量红色文献的数字化存储，但由于红色文献类型复杂多样、数量庞大分散、叙事方式多变等特点，一定程度造成所构建的数据库功能较为单一，存储资源相对独立，较难实现文献内容的语义知识挖掘。现有的红色文献标注体系构建主要集中于单本文献的人工知识标注及常规实体识别，缺乏对红色文献领域标注体系的系统性探索，难以支撑大规模红色文献的自动化标注需求。

鉴于此，课题组前期结合红色文献的内容特征发现，红色文献的内容侧重于对革命人物及革命活动的文本叙述，包括革命事件过程的叙述、革命人物品质的评价、革命人物作品的记载、革命事件环境的描述等。结合 LDA 主题挖掘模型将红色文献的语句划分为事件叙述(标签为 EDS)类、人物评价(标签为 CHA)类、情感表达(标签为 EMO)类、文献摘录(标签为 DOC)类和场景描述(标签为 SCE)类五种语句类别。并依据迁移学习模型与标注体系间的适配性，设计出通用型红色文献的元数据标注体系(见表 1)，包含人物(PER)类、地点(LOC)类、时间(TIM)类、组织(ORG)类、身份(IDE)类、事件(EVE)类六种元数据[31]。

表 1 通用型红色文献的元数据标注体系

元数据名称	标签	含义
人物	PER	人物简称、人物别名和人物名称
地点	LOC	人物籍贯、出生地点、死亡地点、物品地点、事件地点和组织地点
组织	ORG	毕业学校、组织机构和工作单位
时间	TIM	出生日期、死亡日期、物品日期、组织日期和事件日期
事件	EVE	事件名称、死亡方式、文献作品和政策名称
身份	IDE	政治面貌、职务岗位、家庭出身、民族身份和革命信仰

(2)数据来源

本研究构建了革命英烈传记数据集，选取《北平五烈士传》《陈继昌丁香传》《陈君起传》等五十余部革命英烈传记，进行内容特征标注处理，包含事件叙述(标签为 EDS)、人物评价(标签为 CHA)类、情感表达(标签为 EMO)类、文献摘录(标签为 DOC)类和场景描述(标签为 SCE)类五种语句类别；人物(PER)类、地点(LOC)类、时间(TIM)类、组织(ORG)类、身份(IDE)类、事件(EVE)类六种元数据的红色文献语料库。

其中，语句类型采用标签式标注，即按句子为标注单元设定语句类型标签，例如“1927 年 11 月的中央临时政治局扩大会议，通过了《关于第六次全党代表大会之决议》，宣布六大于 1928 年 3 月初至 3 月中旬之间召开。[雨花台烈士传丛书 项英传](标注为 EVE)”。元数据标注方法为 BIESO 序列标注法，即 B 表示元数据的起始位置(Begin)，对应标注为 *B*-元数据类型；I 表示元数据的中间位置(Inside)，对应标注为 *I*-元数据类型；E 表示实体的结束位置(End)，对应标注为 *E*-元数据类型；S 表示单个字为元数据(Single)，对应标注为 S-元数据类型；O 表示非实体字符(Outside)，例如，身份－政治面貌标签的原文标注实例为“共(B-IDE)产(I-IDE)党(I-IDE)员(E-IDE)林(B-PER)祥(I-PER)谦(E-PER)英(O)勇(O)不(O)屈(O)，拒(O)不(O)答(O)应(O)。”

(3)实验数据集

本研究将红色文献语料库按照语句类型进行十等额切分，并以“语句类型＋0X”的方式编号数据集，便于评估基于增量学习的预训练模型微调效果与数据样本间的关联性，即事件叙述类 1 号数据集的编号为“EDS 01”。将人工标注的数据集《朱克靖传》(Data ZKJ)作为常规学习的元数据标注的实验数据集，进行常规学习的红色文献元数据标注评估；未进行人工标注的测试集合《贺瑞麟传》(Data HRL)、《洪灵菲传》(Data HLF)、《孙津川传》(Data SJC)作为迁移学习的元数据标注的数据集，进行迁移学习的红色文献元数据标注评估。

4.2 红色文献元数据标注评价指标和对比模型

红色文献元数据标注任务采用三个常用评价指标，即准确率(Precision，以下简称 P)、召回率(Recall，以下简称 R)，和调和平均值 F_1 值($F_{1-score}$，以下简称 F_1 值)，计算方式详见公式(1)。为更真实准确地评估红色文献元数据标注的效果，降低部分异常实验的噪声，所有模型的评价指标均按十次实验结果的平均值进行评估。

$$\begin{cases} P = \dfrac{TP}{TP + FP} \\ R = \dfrac{TP}{TP + TN} \\ F_1 = \dfrac{2 \times P \times R}{P + R} \end{cases} \tag{1}$$

本研究设计三组基线模型对基于 MRS 6 元数据体系的红色文献元数据标注任务进行性能评估，具体实验方案如下：

(1)基于深度学习的基线模型。以 Keras-GPU 2.6.0 框架构建的深度学习模型，通过 Word2Vec(以下简称 W2C)向量转换后拼接 CNN、BiLSTM、BiGRU、CNN-BiGRU、CNN-BiGRU-CRF 模型。

(2)基于 Transformer 的基线模型。以 TensorFlow-GPU 2.6.0 框架构建的预训练模型，通过 BERT 预训练模型向量转换后拼接 CNN、BiLSTM、BiLSTM-MHA-CRF[31](以下简称 BMC)、BiGRU、CNN-BiGRU、CNN-BiGRU-CRF 和本研究提出的 CBMC 模型。

(3)基于增量学习的基线模型。将增量学习效果最佳的五个模型(BERT-RL-1、BERT-RL-2、BERT-RL-3、BERT-RL-4、BERT-RL-5)后接本研究提出的 CBMC 模型。

4.3 基于增量学习的预训练模型微调性能评估

按照图 2 的增量学习方式对 BERT 预训练模型进行领域微调，具体微调步骤如下：

(1)将按类型切分的五十组微调数据集依次输入 BERT 预训练模型进行识别效果评估，筛选出识别效果最低的微调数据集作为后续增量学习的验证集，实验结果详见表 2。

(2)将表 2 的剩余的四十九组微调数据集依次输入预训练模型进行增量学习训练，将“EDS 07 数据集”按 8∶1∶1 的比例分为训练集、测试集、验证集进行增量学习后模型的效果评估，得出表 3 实验结果，发现该轮增量学习的最佳模型是基于“CHA 09 数据集”微调所得。

表 2　红色文献各切分数据集的元数据识别效果

数据集	F_1 值	数据集	F_1 值	数据集	F_1 值	数据集	F_1 值	数据集	F_1 值
EDS 01	0.8901	CHA 01	0.6014	EMO 01	0.7822	DOC 01	0.8379	SCE 01	0.1576
EDS 02	0.0865	CHA 02	0.7955	EMO 02	0.8407	DOC 02	0.0693	SCE 02	0.8223
EDS 03	0.0359	CHA 03	0.8851	EMO 03	0.0513	DOC 03	0.7949	SCE 03	0.7821
EDS 04	0.5452	CHA 04	0.8254	EMO 04	0.8385	DOC 04	0.8487	SCE 04	0.8568
EDS 05	0.1613	CHA 05	0.2671	EMO 05	0.7185	DOC 05	0.8391	SCE 05	0.7923
EDS 06	0.0045	CHA 06	0.7985	EMO 06	0.0757	DOC 06	0.1711	SCE 06	0.7383
EDS 07	0.0003	CHA 07	0.3964	EMO 07	0.3110	DOC 07	0.8462	SCE 07	0.7472
EDS 08	0.7986	CHA 08	0.7332	EMO 08	0.7531	DOC 08	0.8623	SCE 08	0.8509
EDS 09	0.8093	CHA 09	0.8624	EMO 09	0.7989	DOC 09	0.8264	SCE 09	0.8854
EDS 10	0.7944	CHA 10	0.7194	EMO 10	0.7575	DOC 10	0.0184	SCE 10	0.8004

(3)重复步骤(2)至五十组微调数据集全部完成增量学习训练，将各轮最佳实验结果按 F_1 值绘制图4。选出前五的模型，命名为 BERT-RL-1(第 14 轮结果)、BERT-RL-2(第 13 轮结果)、BERT-RL-3(第 4 轮结果)、BERT-RL-4(第 8 轮结果)、BERT-RL-5(第 12 轮结果)，作为 BERT-RL 的备选模型进行后续实验。

表 3　BERT 预训练模型第一轮增量学习的元数据标注效果

数据集	F_1 值	数据集	F_1 值	数据集	F_1 值	数据集	F_1 值	数据集	F_1 值
EDS 01	0.6376	CHA 01	0.5819	EMO 01	0.5781	DOC 01	0.5918	SCE 01	0.0532
EDS 02	0.0288	CHA 02	0.6171	EMO 02	0.5995	DOC 02	0.1740	SCE 02	0.6210
EDS 03	0.4920	CHA 03	0.6611	EMO 03	0.3718	DOC 03	0.6230	SCE 03	0.6284
EDS 04	0.0794	CHA 04	0.6221	EMO 04	0.6322	DOC 04	0.6369	SCE 04	0.6200
EDS 05	0.3801	CHA 05	0.0362	EMO 05	0.5767	DOC 05	0.6097	SCE 05	0.6174
EDS 06	0.0002	CHA 06	0.6352	EMO 06	0.2886	DOC 06	0.0058	SCE 06	0.6243
EDS 07	——	CHA 07	0.0596	EMO 07	0.3132	DOC 07	0.5880	SCE 07	0.5594
EDS 08	0.6614	CHA 08	0.5480	EMO 08	0.5962	DOC 08	0.6312	SCE 08	0.6467
EDS 09	0.6609	CHA 09	0.6838	EMO 09	0.6353	DOC 09	0.5778	SCE 09	0.5902
EDS 10	0.6084	CHA 10	0.6003	EMO 10	0.0468	DOC 10	0.0019	SCE 10	0.6044

注：由于表 3 中元数据标注效果是基于“EDS 07 数据集”评估所得，故表中“EDS 07 数据集”的值用“——”标识。

通过对表 4 中各备选模型微调数据集的统计分析发现，表现最佳的 BERT-RL-1 模型共计使用了 4 组文献摘录类(DOC)数据，占比最高。该类数据在 MRS 6 元数据表示体系中具有较高的元数据覆盖率，能够为 BERT 预训练模型提供更丰富的语义特征，从而显著提升微调效果。从图 4 中发现，BERT 预训练模型的微调效果与数据规模没有呈现正相关关系，当增量学习轮次超过 14 次后，模型性能出现明显下降趋势，模型效果提升主要依赖于数据类型的多样性，而非单一类型数据的数量积累。因此，预训练模型在跨领域微调过程中，模型准确性的提升更依赖于训练数据集的类型多样性，而非数据集的规模扩张。

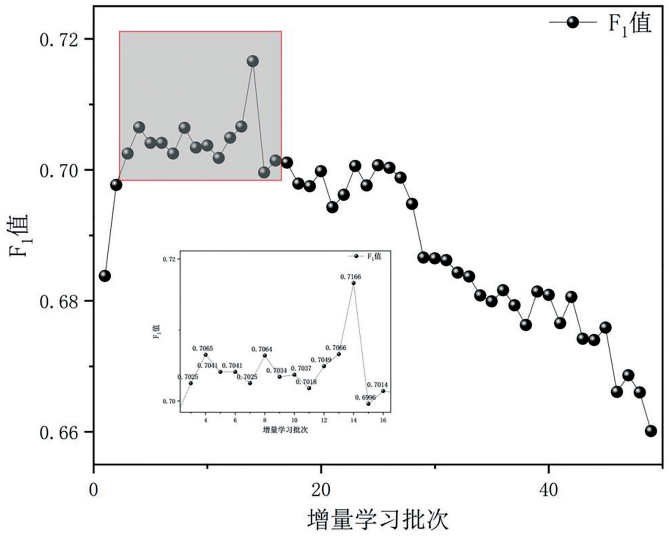

图 4　BERT 预训练模型的增量学习效果

表 4　BERT-RL 备选模型微调时所使用的数据集

模型编号	F_1 值	训练轮次	训练数据集
BERT-RL-1	0.7166	14	CHA 09、EDS 08、DOC 06、DOC 02、EMO 06、SCE 08、EDS 03、DOC 01、CHA 02、CHA 05、SCE 10、EDS 10、SCE 01、DOC 05
BERT-RL-2	0.7066	13	CHA 09、EDS 08、DOC 06、DOC 02、EMO 06、SCE 08、EDS 03、DOC 01、CHA 02、CHA 05、SCE 10、EDS 10、SCE 01
BERT-RL-3	0.7065	4	CHA 09、EDS 08、DOC 06、DOC 02
BERT-RL-4	0.7064	8	CHA 09、EDS 08、DOC 06、DOC 02、EMO 06、SCE 08、EDS 03、DOC 01
BERT-RL-5	0.7049	12	CHA 09、EDS 08、DOC 06、DOC 02、EMO 06、SCE 08、EDS 03、DOC 01、CHA 02、CHA 05、SCE 10、EDS 10

注：增量学习所使用的训练数据仅为模型微调数据集，并非将训练集作为实验的训练集，全程实验数据均为“EDS 07 数据集”按 8∶1∶1 比例切分的训练集、测试集、验证集。

4.4　常规学习的元数据标注评估

本研究以已标注的数据集 Data ZKJ 进行元数据标注实验，各模型的实验结果如表 5 所示。实验结果表明，在常规学习的元数据标注实验中，本研究所设计的 CBMC 模型在融合 BERT-RL-1 模型后，显著提升红色文献的元数据标注效果，且模型在各类型元数据标注任务中均具有较好的适配性。

实验结果表明，由于事件类（EVE）和时间类（TIM）元数据表达形式具有多样化的特点，传统 BERT 预训练模型在词边界识别任务中较难准确判断元数据的词边界。采用增量学习策略微调的 BERT-RL-1 模型在事件类（EVE）元数据识别任务中，较基线模型 BERT-CBMC 的 F_1 值提升 11.83 个百分点；在时间类（TIM）元数据识别任务中，较基线模型 BERT-CBMC 提升 11.00 个百分点。此外，该模型在其他类型元数据的识别中展现出良好的知识迁移能力。BERT-RL-1-CBMC 模型在 MRS 6 元数据表示体系上的 F_1 值达 0.8914，较基线模型 BERT-CBMC 提升 5 个百分点。上述实证结果不仅验证了增量学习在领域适配性模型优化中的有效性，更验证了大语言模型微调范式在红色文献领域的适应性。

4.5　迁移学习的元数据标注评估

为证明 BERT-RL-1 和 CBMC 模型在红色文献领域的元数据标注的迁移性，本研究在表 5 的实验结果基础上，选择 BERT＋BMC 模型、BERT＋CBMC 模型、基于增量学习的基线模型，尝试对未标注的 Data HRL 进行元数据的自动化标注实验，元数据标注的实验结果如图 5 所示。

表 5　数据集 Data ZKJ 的元数据标注效果评估

模型	元数据标签						组合		
	TIM	EVE	ORG	PER	IDE	LOC	P	R	F_1 值
W2C	0.1152	0.0200	0.1824	0.6859	0.0482	0.2924	0.6760	0.2761	0.3884
W2C+CNN	0.6712	0.2016	0.4247	0.8028	0.4076	0.5661	0.8621	0.4413	0.5816
W2C+BiLSTM	0.4815	0.1789	0.5416	0.8295	0.4824	0.5080	0.8604	0.4775	0.6125
W2C+BiGRU	0.4864	0.2189	0.5714	0.8366	0.5085	0.4688	0.8510	0.4951	0.6260
W2C+CNN-BiGRU	0.6883	0.3051	0.5799	0.8456	0.5119	0.5549	0.8836	0.5176	0.6528
W2C+CNN-BiGRU-CRF	0.7037	0.2918	0.6036	0.8430	0.5618	0.5831	0.7663	0.5741	0.6564
BERT	0.8564	0.6680	0.7934	0.9256	0.7774	0.8107	0.9139	0.7580	0.8285
BERT-CNN	0.8719	0.6090	0.7762	0.9089	0.7994	0.7304	0.8925	0.7363	0.8067
BERT-BiLSTM	0.8821	0.6513	0.7974	0.9181	0.8125	0.8281	0.8910	0.7804	0.8320
BERT-BMC[31]	0.9129	0.6461	0.7960	0.9277	0.8258	0.8294	0.9013	0.7823	0.8375
BERT-BiGRU	0.8512	0.6212	0.8068	0.9282	0.8136	0.8463	0.8942	0.7837	0.8352
BERT-CNN-BiGRU	0.8564	0.6644	0.7796	0.9293	0.8225	0.8355	0.9221	0.7665	0.8370
BERT-CNN-BiGRU-CRF	0.8621	0.6648	0.7954	0.9277	0.8339	0.8212	0.9036	0.7843	0.8396
BERT-CBMC	0.8270	0.6745	0.8123	0.9303	0.8167	0.8188	0.9095	0.7830	0.8414
BERT-RL-1-CBMC	0.9370	0.7928	0.8674	0.9413	0.8750	0.8681	0.9622	0.8303	0.8914
BERT-RL-2-CBMC	0.9589	0.7919	0.8658	0.9402	0.8720	0.8567	0.9603	0.8272	0.8888
BERT-RL-3-CBMC	0.9351	0.8010	0.8655	0.9346	0.8698	0.8596	0.9614	0.8235	0.8871
BERT-RL-4-CBMC	0.9212	0.7690	0.8551	0.9372	0.8544	0.8490	0.9588	0.8108	0.8786
BERT-RL-5-CBMC	0.8505	0.7551	0.8444	0.9296	0.8383	0.8296	0.9373	0.8021	0.8644

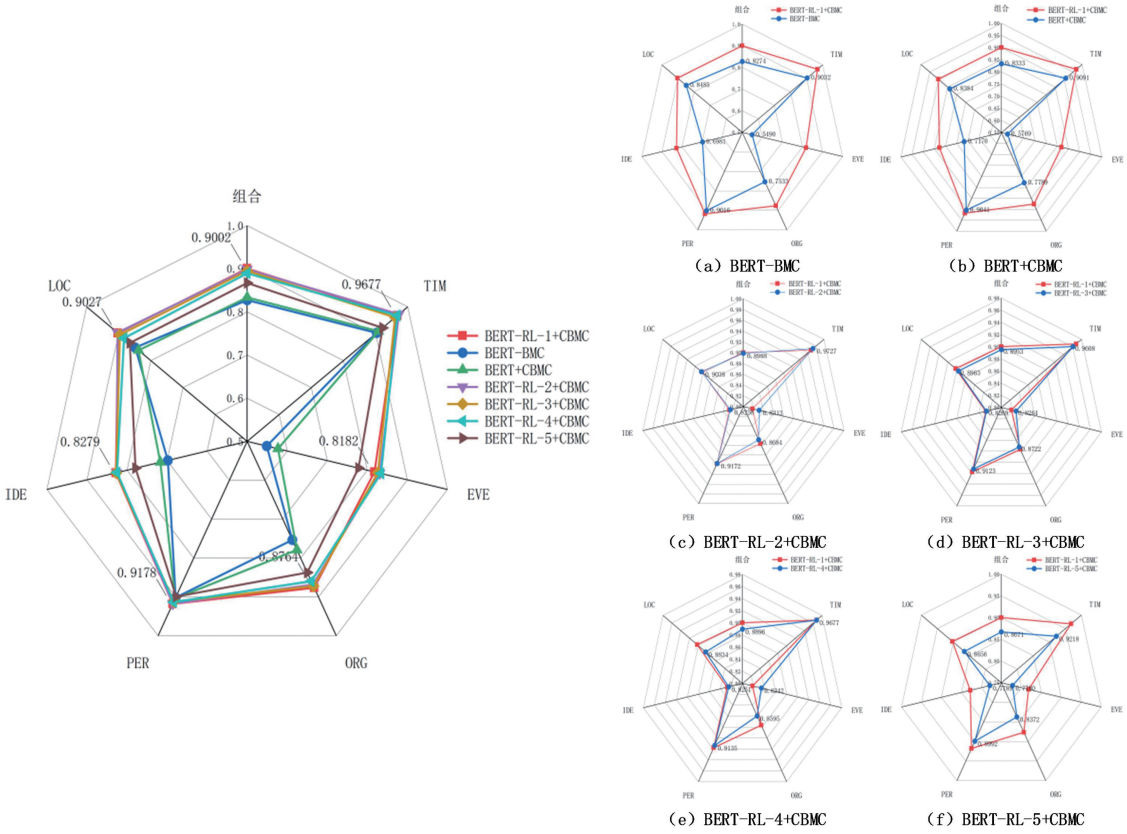

图 5　各模型在 Data HRL 的元数据自动化标注 F_1 值对比图

图 5 左侧总图可以发现，BERT-RL-5＋CBMC 的综合性能最优，F_1 值为 0.9002，能较为准确地实现对红色文献的迁移标注。图 5 右侧的不同模型与 BERT-RL-5＋CBMC 的效果对比可以发现，如图 5-(a)和图 5-(b)所示，增量学习的微调方法能有效提升预训练模型在处理红色文献时的准确性，在组织

(ORG)、事件(EVE)和身份(IDE)类等元数据的标注效果均获得显著提升。其原因是通过增量学习微调后的预训练模型提升了对红色文献领域组织(ORG)、事件(EVE)和身份(IDE)类元数据词边界的界定能力。例如“1930 年邓振询成为兴国县工人运动领袖，1932 年任江西省职工联合会委员长，他“为了大家舍小家”，全身心地扑在革命工作上”的预训练模型 BERT 和基于增量学习的预训练模型 BERT-RL-1 的标注效果对比详见表 6。

表 6　不同模型的标注效果对比

标注方式	标注结果								
原文	1	9	3	0	年	邓	振	询	成
BERT	B-TIM	I-TIM	I-TIM	I-TIM	E-TIM	B-PER	I-PER	E-PER	O
BERT-RL-1	B-TIM	I-TIM	I-TIM	I-TIM	E-TIM	B-PER	I-PER	E-PER	O
原文	为	兴	国	县	工	人	运	动	领
BERT	O	B-LOC	I-LOC	E-LOC	B-EVE	I-EVE	I-EVE	E-EVE	B-IDE
BERT-RL-1	O	B-IDE	I-IDE	I-IDE	I-IDE	I-IDE	I-IDE	I-IDE	I-IDE
原文	袖	，	1	9	3	2	年	任	江
BERT	E-IDE		B-TIM	I-TIM	I-TIM	I-TIM	E-TIM	O	B-ORG
BERT-RL-1	E-IDE		B-TIM	I-TIM	I-TIM	I-TIM	E-TIM	O	B-IDE
原文	西	省	职	工	联	合	会	委	员
BERT	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG	E-ORG	B-IDE	I-IDE
BERT-RL-1	I-IDE	I-IDE	I-IDE	I-IDE	I-IDE	I-IDE	I-IDE	I-IDE	I-IDE
原文	长	，	他	“	为	了	大	家	舍
BERT	E-IDE		O		O	O	O	O	O
BERT-RL-1	E-IDE		O		O	O	O	O	O
原文	小	家	”	，	全	身	心	扑	在
BERT	O	O			O	O	O	O	O
BERT-RL-1	O	O			O	O	O	O	O
原文	革	命	工	作	上	。			
BERT	B-EVE	E-EVE	O	O	O				
BERT-RL-1	O	O	O	O	O				

数据来源：张新贵. 雨花台烈士传丛书 邓振询传[M]. 南京：江苏人民出版社，2016.

由表 6 中的标注结果可见：在处理“江西省职工联合会委员长”文本时，基线模型 BERT 预训练模型错误地将其切分为江西省职工联合会(ORG)和委员长(IDE)两个词语，而实际应标注为江西省职工联合会委员长(IDE)。值得关注的是，经过增量学习优化的 BERT-RL 预训练模型能够准确识别此类不应切分的复合实体，提升了模型整体识别效果。在进一步对 BERT 预训练模型和 BERT-RL 系列预训练模型处理结果中存在差异的句子集中分析时发现，差异样本均存在“触发词＋元数据”的结构特征(如“任(触发词)＋身份类(IDE)元数据”“成为(触发词)＋身份类(IDE)元数据”等模式)。在后续研究中，可尝试构建红色文献领域触发词词表，进一步提升红色文献的元数据标注效果。

为进一步证明 BERT-RL-1 和 CBMC 模型在红色文献领域的自动化标注的泛化能力，本研究将上述模型对未标注的数据集 Data HLF 和 Data SJC 进行元数据标注实验，元数据标注结果如表 7 所示。

表 7　数据集 Data HLF 和 Data SJC 的元数据标注效果评估

模型	Data HLF			Data SJC		
	P	R	F_1 值	P	R	F_1 值
BERT-BMC[31]	0.8643	0.8864	0.8752	0.9172	0.7779	0.8418
BERT-CBMC	0.9004	0.8938	0.8971	0.9180	0.7867	0.8473
BERT-RL-1-CBMC	0.9711	0.9184	0.9440	0.9763	0.8327	0.8988
BERT-RL-2-CBMC	0.9696	0.9195	0.9439	0.9789	0.8304	0.8985
BERT-RL-3-CBMC	0.9685	0.9187	0.9429	0.9740	0.8241	0.8928
BERT-RL-4-CBMC	0.9694	0.9117	0.9397	0.9586	0.8185	0.8830
BERT-RL-5-CBMC	0.9715	0.9065	0.9379	0.9469	0.7885	0.8604

实验结果表明 BERT-RL-1 和 CBMC 模型在其他红色文献的元数据标注实验结果中的效果同样最优，具有较好的普适性。因此，本研究选择 BERT-RL-1 模型作为 BERT-RL 模型的参数继承模型。

4.6　红色文献元数据标注消融实验

为进一步证明 BERT-RL 和 CBMC 模型的延展性，本研究在数据集 Data SJC 的基础上开展详细的红色文献元数据标注消融实验。结合前文的实验结果，本研究选择 BERT 预训练模型和 BERT-RL 预训练模型结合深度学习模型进行消融实验，元数据标注的实验结果如表 8 所示。实验结果表明，相较于 BERT 预训练模型，BERT-RL 预训练模型具有更好的鲁棒性与延展性。

表 8　数据集 Data SJC 的元数据标注消融实验

模型	BERT			BERT-RL		
	P	R	F_1 值	P	R	F_1 值
本身	0.8849	0.6851	0.7723	0.8937	0.7747	0.8286
＋CNN	0.9004	0.6790	0.7742	0.9656	0.7606	0.8509
＋BiLSTM	0.7614	0.5245	0.6211	0.9532	0.7871	0.8622
＋BMC[31]	0.9172	0.7779	0.8418	0.9580	0.8267	0.8873
＋BiGRU	0.9154	0.7742	0.8389	0.9650	0.7883	0.8677
＋CNN-BiGRU	0.9272	0.7666	0.8393	0.9688	0.8146	0.8850
＋CNN-BiGRU-CRF	0.9215	0.7804	0.8451	0.9798	0.8250	0.8958
＋CBMC	0.9180	0.7867	0.8473	0.9763	0.8327	0.8988

4.7　基于 BERT-RL-CBMC 的“雨花英烈人物传记语料库”构建

本研究应用 BERT-RL-CBMC 模型对红色文献进行大规模标注，构建了“雨花英烈人物传记体语料库”，具体构建过程详见图 6。

“雨花英烈人物传记语料库”共涵盖 95637 个人物（PER）类元数据、60582 个地点（LOC）类元数据、83066 个组织（ORG）类元数据、18630 个时间（TIM）类元数据、30733 个事件（EVE）类元数据、64600 个身份（IDE）类元数据，具体数据实例详见表 9。

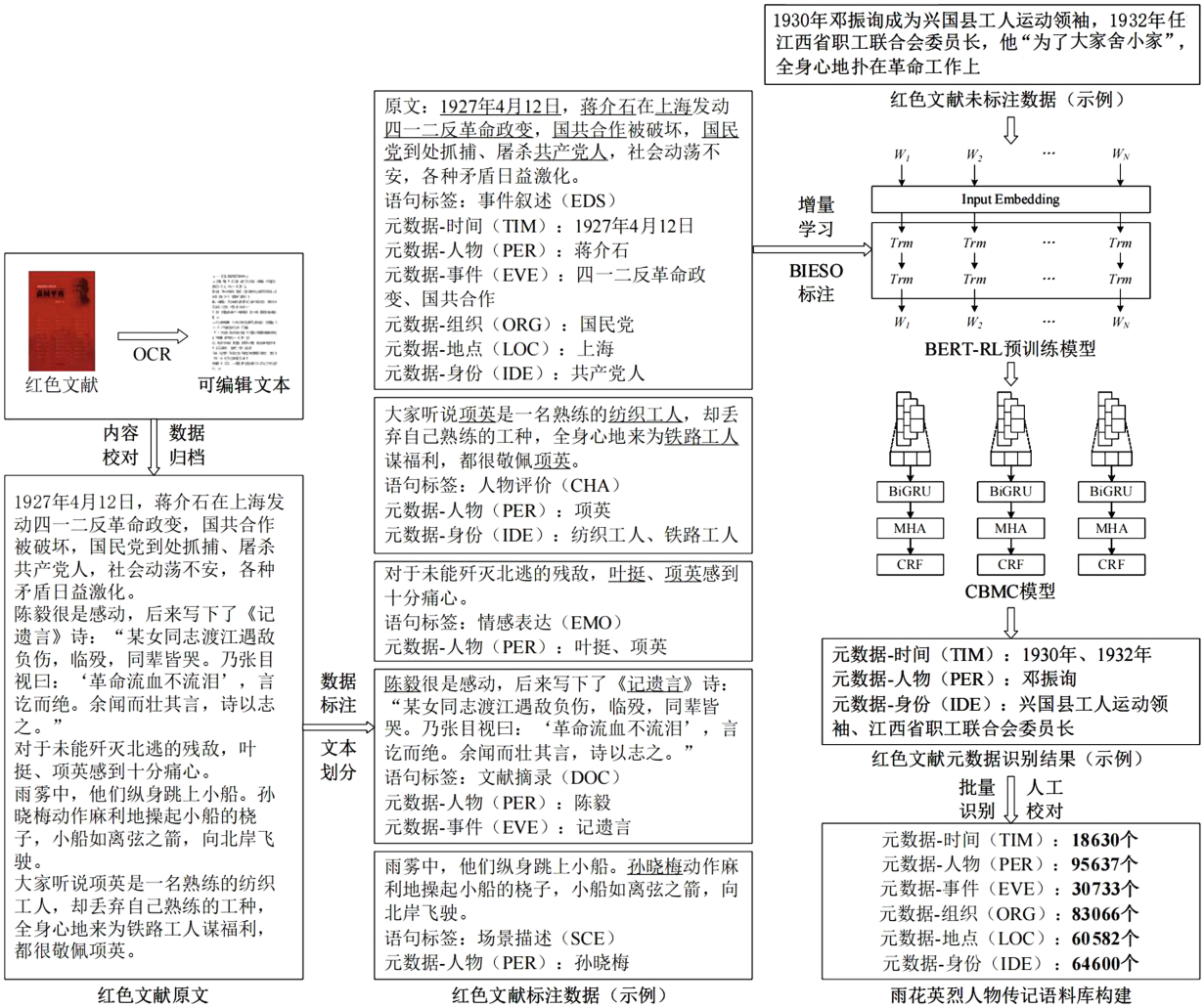

图 6　基于 BERT-RL-CBMC 模型的“雨花英烈人物传记语料库”构建示例图

表 9　“雨花英烈人物传记语料库”实例

标签	数量(个)	示例		
		元数据	原文	文献来源
PER	95637	朱克靖	朱克靖是湖南醴陵人，1895 年 10 月 11 日出生在醴陵北乡株树下村的朱家老屋，本名宏夏，字竹懿，亦作竹怡，克靖是他的号，当时社会习俗以号行世	《雨花台烈士传丛书 朱克靖传》
LOC	60582	武汉	在热潮推动下，一九一七年下半年，武汉地区诞生了第一个进步团体——互助社	《雨花台烈士传丛书 恽代英传》
ORG	83066	景山小学	孙晓梅去找景山小学的老师们，去找孙氏族亲孙京良，请求他们帮忙寻一份工作，她要自食其力	《雨花台烈士传丛书 孙晓梅传》
TIM	18630	1913 年 3 月	1913 年 3 月，邓中夏以优异成绩考入郴郡六城联立中学(1914 年改为湖南省第七联合中学)	《雨花台烈士传丛书 邓中夏传》
EVE	30733	武昌起义	1911 年 10 月 10 日，湖北革命团体文学社、共进社在同盟会的推动下，以湖北新军为主力发动了武昌起义	《雨花台烈士传丛书 项英传》
IDE	64600	老贡生	6 岁时，金佛庄进入老贡生金洪锦的私塾启蒙	《雨花台烈士传丛书 金佛庄传》

数据来源：课题组收集的 55 部红色文献。

5 结语

本研究在课题组已标注的红色文献数据基础上，通过增量学习的方式微调出能够高精度捕获红色文献文本特征的 BERT-RL 预训练模型，并针对 BERT-RL 预训练模型缺乏对红色文献上下文语义信息学习的问题，构建 CBMC 的标注模型。实验结果表明，本研究提出的融合 BERT-RL 和 CBMC 的红色文献元数据标注模型能精准地捕获元数据的词边界特征，提升向量转换的准确性，有效提升模型整体识别效果。此外，通过对其他红色文献的迁移标注测试，以及对模型的消融对比试验，有效证明 BERT-RL 预训练模型具有较好的泛化能力和鲁棒性，促进红色文献元数据标注研究的发展。据此，通过增量学习方式微调出的 BERT-RL 预训练模型为特定语境或专业知识场景下大语言模型应用提供较好的理论基础，可视为大语言模型在跨领域微调技术的推广。在未来的研究中，本研究构建的标注模型将继续应用于红色文献知识库构建中，基于该知识库将有效解决红色文献知识与大语言模型内部参数化知识深度融合问题，更好地解决大语言模型在红色文献知识服务中的可靠性问题。

参考文献

1 马静. 赓续与弘扬：革命历史文献整理研究的实践与思考[J]. 中国图书馆学报，2023，49(4)：34－41.
2 夏翠娟. 数字人文的理论沿革和范式转向：从“人文研究的数字方法”到“后数字时代的人文研究”[J/OL]. 中国图书馆学报，1－24[2024－10－16]. http://kns.cnki.net/kcms/detail/11.2746.G2.20240506.1825.002.html.
3 马静. 探索与求新：革命文献与民国时期文献保护计划的实践与思考[J]. 中国图书馆学报，2022，48(4)：79－89.
4 张钦彤，王昱超，王鹤羲，等. 大语言模型微调技术的研究综述[J]. 计算机工程与应用，2024，60(17)：17－33.
5 Yosinski J，Clune J，Bengio Y，et al. How transferable are features in deep neural networks? [J]. Advances in Neural Information Processing Systems，2014(27)：3320－3328.
6 刘倩倩，夏翠娟，朱武信. 红色文化传承视域下的红色文献服务平台建设实践与思考[J]. 信息资源管理学报，2021，11(4)：17－24，32，16.
7 徐孝娟，孙爱华，史如菊，等. 数字人文视角下皖西红色文献知识本体及其应用[J]. 图书馆论坛，2023，43(10)：139－151.
8 陈忻，房小可，孙鸣蕾. 社会记忆再生产：北京香山红色档案编研成果的细粒度挖掘研究[J]. 山西档案，2021(1)：80－87，79.
9 毕崇武，张晓琪，周静虹. 数据态环境下红色文献的语义化表示方法研究[J]. 图书馆学研究，2024(3)：59－70.
10 付靖宜，李姗姗，项欣溢，等. 新四军苏浙军区多模态红色文献资源知识聚合模式研究[J]. 档案学研究，2022(4)：24－31.
11 何琳，武帅，叶雅慧，等. 基于轻量级本体的雨花英烈群体特征构建与分析研究[J]. 农业图书情报学报，2022，34(12)：33－44.
12 王雨生. 数字人文视域下红色文献资源知识图谱构建研究[D]. 长春：长春师范大学，2024.
13 淳鑫，冯玲，李航. 面向红色文化的命名实体识别研究[J]. 无线电通信技术，2023，49(4)：622－628.
14 吴艳飞，张强，周树斌，等. 数字人文视角下的大别山革命文物知识图谱构建[J]. 知识管理论坛，2024，9(2)：148－160.
15 裴炳森，李欣，蒋章涛，等. 基于大语言模型的公安专业小样本知识抽取方法研究[J]. 计算机科学与探索，2024，18(10)：2630－2642.
16 王润周，张新生，王明虎. 融合动态掩码注意力与多教师多特征知识蒸馏的文本分类[J]. 中文信息学报，2024，38(3)：113－129.
17 胡昊天，邓三鸿，王东波，等. 情报学视角下的预训练语言模型研究进展[J]. 图书情报工作，2024，68(3)：130－150.
18 Tanwisuth K，Zhang S，Zheng H，et al. POUF：prompt－oriented unsupervised fine－tuning for large pre－trained models[C]//International Conference on Machine Learning. PMLR，2023：33816－33832.
19 Liu H，Tam D，Muqeeth M，et al. Few－shot parameter－efficient fine－tuning is better and cheaper than in－context learning[J]. Advances in Neural Information Processing Systems，2022，35：1950－1965.
20 林令德，刘纳，王正安. Adapter 与 Prompt Tuning 微调方法研究综述[J]. 计算机工程与应用，2023，59(2)：12－21.
21 Radford A，Narasimhan K，Salimans T，et al. Improving language understanding by generative pre-training[R]. OpenAI，2018.
22 Veyseh A P B，Lai V，Dernoncourt F，et al. Unleash GPT－2 power for event detection[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1：Long Papers). 2021：6271－6282.
23 Ulfah U，Arifudin O. Analisisteori taksonomi bloom pada pendidikan di Indonesia[J]. Jurnal Al－Amar：Ekonomi Syariah，Perbankan Syariah，Agama Islam，Manajemen Dan Pendidikan，2023，4(1)：13－22.
24 Biswas S S. Role of ChatGPT in public health[J]. Annals of Biomedical Engineering，2023，51(5)：868－869.
25 Devlin J，Chang M W，Lee K，et al. BERT：pre－training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Stroudsburg，PA：Association for Computational Linguistics，2019：4171－4186.
26 Adoma A F，Henry N M，Chen W. Comparative analyses of bert，roberta，distilbert，and xlnet for text-based emotion recognition[C]//2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE，2020：117－121.
27 曹欢欢，李玥琪. 基于 ALBERT 和 BiLSTM－Attention 的高校网络舆情监测模型研究[J/OL]. 情报科学，1－14[2024－07－10]. http://kns.cnki.net/kcms/detail/22.1264.g2.20240604.

1047.004.html.

28 Shin H C，Zhang Y，Bakhturina E，et al. BioMegatron：larger biomedical domain language model[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020：4700－4706.

29 Maheshwari H，Singh B，Varma V. SciBERT sentence representation for citation context classification[C]//Proceedings of the Second Workshop on Scholarly Document Processing. 2021：130－133.

30 Fan W，Wang D，Huang S. Automatic sentence segmentation for classical Chinese：the Spring and Autumn Annals as an example[J]. Digital Scholarship in the Humanities，2023，38(3)：1067－1077.

31 武帅，何琳，杨海龄，等.迁移学习视角下红色文献元数据表示体系构建探究[J].情报资料工作，2024，45(6)：84－92.

作者贡献说明：
武帅：论文思想的具体实现，论文撰写
何琳：提出论文的研究思路，论文修订
杨秀璋：参与实验的修改完善，代码指导
孙邵丹：数据清洗和方案设计，论文修订
桂思思：参与论文的修改完善，论文修订

作者单位：武帅、何琳、孙绍丹、桂思思，南京农业大学信息管理学院，江苏南京，210095
杨秀璋，贵州大学公共大数据国家重点实验室，贵州贵阳，550025

收稿日期：2025 年 1 月 13 日
修回日期：2025 年 4 月 1 日

（责任编辑：关志英）

Research on Metadata Annotation of Red Literature Based on BERT-RL

WU Shuai　HE Lin　YANG Xiuzhang　SUN Shaodan　GUI Sisi

Abstract: Red Literature is generally characterized by its large volume, dispersion, and complexity and variety of types, making it more difficult to achieve efficient application of large language models based on zero-based training. How to use a small amount of annotated data to fine-tune a large language model for accurate, efficient and scalable Automated Labeling of metadata in the Red Literature domain is one of the key problems that need to be solved nowadays. The aim of this study is to explore an efficient fine-tuning strategy for large language models to enhance the model's adaptability, contextual comprehension and annotation accuracy in the domain of semantically rich but resource-scarce texts. Methodologically, the study designed a large language model fine-tuning method based on Incremental Learning and combined it with deep learning techniques to achieve Automated Labeling of Red Literature Metadata Representation System. Firstly, according to the content characteristics of Red Literature, it was categorized into five types of textual datasets: event narratives, character evaluations, emotional expressions, literature excerpts and scene descriptions. Second, an Incremental Learning strategy was used to train the BERT pre-training model in multiple rounds of iterations, and the parameter combinations with the best results are screened and inherited to the BERT-RL model. Finally, based on BERT-RL, this study proposed a composite annotation model, BERT-RL-CBMC, which incorporated Convolutional Neural Network (CNN), Bi-directional Gated Recurrent Unit (BiGRU), and Multi-Head Attention(MHA) with Conditional Random Field (CRF), and was designed to capture the short-range lexical features and long-range semantic dependencies in Red Literature sequence texts. By conducting multi-group comparative experiments in Red Literature, it is shown that the BERT-RL-CBMC model significantly outperforms the traditional baseline model in boundary identification of complex metadata categories such as Organization (ORG), Event (EVE) and Identity (IDE), which enhances the model annotation accuracy; and exhibits excellent generalization ability in transfer learning tasks, and the ablation experiments validate the complementary roles of the modules in the architecture . This study validates the effectiveness of Incremental Learning strategies in domain adaptation of large language models and provides a theoretical basis for the application of large language models in specific knowledge scenarios. Future research will continue to apply the model to construct a Red Literature knowledge base, solve the problem of deep integration of domain knowledge and model parameterized knowledge, and improve the reliability of Red Literature knowledge services.

Keywords: Incremental Learning; Red Literature; Metadata Representation System; Automated Labeling; BERT-RL