融合旋转式位置编码与图递归检索方法的书院事件抽取研究

喻雪寒　何琳*

摘要　书院是我国古代独特的教育机构，而《中国书院辞典》作为记载书院的重要资料，收纳自唐代至清代全国有史可考的书院多达1600余所。为全面、系统地整理与提取有效数据，文章在对事件抽取各类模式与方法综述的基础上，探索出综合旋转式位置编码与图递归检索的方法以抽取书院的事件信息：利用RoFormerV2模型对绝对位置进行编码，使每个向量附带相对位置信息，之后借助全局归一化思想通过嵌套实体识别模型GlobalPointer和完全子图搜索方式递归查找事件类型与论元。在《中国书院辞典》上进行的实验表明，该方法能有效融合向量的位置和语义信息并对论元间的关联性进行建模，克服了长文本引发的信息缺失与事件论元的嵌套问题，并具备良好的外推性。

关键词　中国书院辞典　事件抽取　RoFormerV2　GlobalPointer　图递归检索
分类号　K061　TP391.1
DOI　10.16603/j.issn1002－1027.2025.02.006
引用本文格式　喻雪寒，何琳.融合旋转式位置编码与图递归检索方法的书院事件抽取研究[J].大学图书馆学报，2025，43(2)：50－65.

书院作为东亚儒家文明的典型形态，也作为我国文化中一种独具特色的教育模式，以教书、藏书、著书、刻书为主体，自唐代至晚清已传承千余年。由季啸风所著的《中国书院辞典》则记载了千年以来我国书院大到初创建立、改制废除，小到学制学规、著书刻书等一系列重要史料[1]，若能高效准确提取出辞典里关键的知识元素，如地点、朝代、人物、官职、行为等信息，将不仅可以展现儒家文化的区域性差异，同时还可揭示其构建主体的阶级变化。

当前在信息抽取层面，依据信息要点的不同可分为实体、关系、事件，在此基础上信息抽取的子任务进一步被划分为命名实体识别、关系抽取、事件抽取。文章选用的事件抽取任务本质属于多元关系抽取，是将文本根据语义描述差异归纳为各类事件并进行细粒度识别，帮助学者快速识别文本数据中的重要信息，归纳整理以便用于后期事理图谱及信息检索系统的构建。

事件抽取作为自然语言处理领域的重要分支之一，最早可追溯至上世纪90年代，美国国家标准与技术研究所召开的自动内容提取会议(Automatic Content Extraction，ACE)将事件定义为：在特定时间、地点中，由一个或多个角色引发的一个或多个动作造成了事物状态的改变，从而推动了该任务的发展①。现阶段，大多数的事件抽取任务建立在句子级基础上，常见的BERT[2]、RoBERTa[3]、ELECTRA[4]均使用了绝对位置编码，使文本的最大输入长度限制为512个字符。在对《中国书院辞典》这一文本集进行梳理后，发现该类语料存在两大特征：一方面，词条以书院为单位，部分词条字数超出常规预训练模型的文本输入要求，而超出最大长度的粗暴截断致使句子级的预训练模型不能准确对其建模；另一方面，语料存在不同事件类型共用相同触发词的现象，即一个触发词可表征多个事件类型，而传统的事件抽取任务将触发词识别视作序列标注任务，忽略了触发词和事件论元间的相互关联。

为了解决上述问题，文章融合旋转式位置编码与图递归检索方法，并基于《中国书院辞典》的内容

* 通讯作者：何琳，邮箱：helin@njau.edu.cn。

① Linguistic Data Consortium. ACE(Automatic Content Extraction) english annotation guidelines for events[EB/OL].[2025－01－10]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf.

构建一种可处理长文本的事件抽取方法，具体贡献如下：

(1)针对绝对位置编码的最长输入截断现象，使用基于旋转式位置编码的 RoFormerV2 模型，利用旋转矩阵对绝对位置进行编码，使每个向量附带相对位置信息，令其具有更好的外推性。

(2)为了解决事件论元嵌套与论元间的关联性问题，接入了图递归检索模块 GPLinker。该模块通过一个嵌套实体识别模型 GlobalPointer[5]将识别得到的(事件类型，触发词，具体触发词)和(事件类型，论元角色，论元)作为完全图的相邻节点，用递归检索策略进行抽取，只有相同事件类型的论元和触发词才会被关联上，解决了触发词的误识别问题。

(3)基于事件抽取结果，利用时空统计法梳理明清时期书院创办的空间分布特征和建设力量差异，论证明清两代书院创办呈现的特征分析。

1 相关研究

1.1 事件抽取模式

事件抽取指在确定的语句结构中，从句子或篇章文本中识别出符合要求的事件信息，具体可细分为触发词识别、事件类型分类、论元提取、论元角色分类四个子任务，前两者可称作事件检测，后两者则合并为元素识别。

一般来说，当子任务先后串行提取时被视作是管道型抽取，即先识别触发词并判断相应的事件类型，再检测论元确定其论元角色。比较典型的有陈(Chen)等人的动态多池化卷积神经网络(DMCNN)模型，在事件检测阶段将句子经过卷积得到的特征分段进行池化，以捕获句子不同部位的突出特征，然后在元素识别阶段根据触发词与事件论元的位置将各部分池化的结果拼接构成句级特征从而抽取[6]；郭鑫等人提出三阶段管道式方法，先用无监督方式对事件类型分类，接着进行事件句提取，最后利用 BiLSTM—CRF 模型识别并拼接事件论元[7]；王(Wang)等人构建基于问答的篇章级核心事件抽取模型，先检测事件类型，再利用问答模式的 BTBiLSTM 提取事件论元，最后融合并选择得分最高的核心事件[8]。

但是，上述管道型抽取的弊端较为明显，下游任务对上游抽取结果的依赖性决定了错误容易产生级联，一旦事件检测失误，对应的论元识别效果也将大大降低。为了解决这类问题，进而衍生了联合型的抽取方式，该模式利用联合提取算法同时预测事件类型与具体论元，克服了级联错误的发生。李(Li)等人提出了结合局部和全局特征的结构化预测框架，可同时提取触发词和论元[9]；阮(Nguyen)等人提出基于双向递归神经网络的联合框架，引入内存矩阵，以此捕获论元角色和触发词的依赖关系[10]；葛军伟等人利用 BiLSTM 模型提取段落特征，采用自注意力机制获取上下文交互信息，融合文档序列更新语义表示，最后采用序列标注提取论元和事件类型[11]。

1.2 事件抽取方法

事件抽取最初始于基于模式匹配的方法，抽取模板是由专家融合专业背景知识和不同的句法、语法特征来人工设计。随着算法技术的发展和算力水平的提升，相继涌现出基于机器学习、深度学习的方式。

基于机器学习的事件抽取本质是将抽取转化为分类问题，运用支持向量机、条件随机场、隐马尔可夫模型等算法构建分类器进行事件分类与论元识别。贝塔德(Bethard)和马丁(Martin)结合形态句法特征和支持向量机的方法构建事件检测系统[12]；洛朗(Llorens)等人使用条件随机场算法，利用各类句法和语义角色特征进行事件识别与分类[13]；博罗什(Boros)等人通过无监督学习获得词向量特征表示，使用决策树分类抽取具体的事件论元[14]。

机器学习弱于语义特征的学习，不擅长处理复杂的语义关系，而深度学习的方式恰好弥补了这一缺陷。张(Zhang)等人充分利用跳窗卷积神经网络提取全局的结构化特征，再通过波束搜索寻找句子，提高触发词识别的准确性[15]；段(Duan)等人的文档级循环神经网络(DLRNN)模型，通过分布式向量的文档表示来提取跨句线索，连接文档向量和词嵌入特征并将其作为 BiLSTM 模型的输入[16]；薛颂东等人采用多粒度阅读器实现多层次语义编码，通过图注意力网络获取实体对的全局与局部关系，构造剪枝完全图为触发器捕捉事件和论元[17]。

在深度学习领域，预训练模型也已成为一种强有力的技术手段。它的出现源于现实中缺乏足够的标注数据，而深度学习模型往往需要大量数据来训练从而避免过拟合问题，因此使用预训练模型汲取在大规模数据集上学习到的固定特征来帮助解决小样本难题。其中，位置编码记录了字词在文本中的顺序，使句子变为具备前因后果的字词序列，预训练模型里常见的编码方式有绝对位置编码、相对位置

编码。绝对位置编码基于位置嵌入，对每个位置都分配了一个唯一的位置向量，这种向量固定的编码方式对短序列较为友好，但不擅长处理超过模型训练长度的长序列文本，田三川与张虎在事件抽取研究中用到的 BERT 和 RoBERTa 都属于绝对位置编码[18-19]。相对位置编码基于相对位置，给每个位置赋予一个表示该位置与其他位置相对距离的偏移量，这类编码方式可用于处理长序列，并在上下文保持一致性，较为典型的有 XLNET[20]、T5 模型[21]。然而，当前的相对位置编码大多基于注意力矩阵进行操作，而本文所用到的旋转式位置编码 RoFormerV2 事先不用计算注意力矩阵，而是以内积形式使向量附带相对位置信息，该编码方式存在两大优点：一是擅长处理长文本的语义信息，二则是捕捉长句中的相似结构和重复模式。本文所使用的《中国书院辞典》存在部分词条超出常规预训练最大文本长度的情况，且辞典本身是将知识信息经过系统化整理和归纳后的产物，存在一定的制式性，适用于 RoFormerV2 捕捉对称性数据的特点。

2 研究方法

2.1 模型架构

文章使用的融合旋转式位置编码与图递归检索方法的模型架构如图 1 所示，该模型包括基于旋转式位置编码的预训练模型 RoFormerV2、由 GlobalPointer 和完全图组成的图递归检索模块 GPLinker 两部分，用以联合抽取事件类型、触发词、论元角色及论元。

当文本序列输入到模型中时，预训练模型 RoFormerV2 依据特有的旋转位置嵌入(Rotary Position Embedding，RoPE)方法，将 transformer 里位置为 m、n 的 query、key(简称 q、k)向量分别乘以旋转矩阵得到变换后的 q、k 向量，而输入的文本序列也将充分利用向量的相对位置信息，进一步从语义上获取语料库的上下文特征，形成书院数据的向量表示。

上述向量表示输入至图递归检索模块 GPLinker 时，会先被 GlobalPointer 这一嵌套实体识别模型进行首尾标注，重点标注事件类型、论元角色和触发词。GlobalPointer 在记录位置的同时也将“(事件类型，论元角色)、(事件类型，触发词)”各自组合形成实体加以识别，识别后的(事件类型，论元角色/触发词，论元/具体触发词)会被送进完全图作为图中的节点，相同事件类型成为相邻节点，使完全子图搜索转化为对同一“论元/具体触发词”所对应的不同事件的递归检索。

2.2 旋转式位置编码

RoFormerV2 模型建立在 RoFormer 模型[22]基础上，出发点是用绝对位置与旋转矩阵的乘积进行编码，以达到用“绝对位置编码方式实现相对位置编码”的效果。为结合相对位置信息，RoFormer 模型先给 transformer 的 q、k 添加绝对位置信息 m、n 将其转为 q_m 和 k_n，用函数 g 表示 q_m 和 k_n 的内积，其中 x_m 和 x_n 是词嵌入向量，$m-n$ 是它们的相对位置信息，最终目标是找到一种等效机制以符合公式(1)的函数关系：

$$(f_q(x_m,m),f_k(x_n,n))=g(x_m,x_n,m-n) \quad (1)$$

当 $x_i \in R^d$，由于内积满足线性叠加性，我们将 d 维空间拆分成 d/2 个子空间并以内积形式进行线性拼接，向量 x 和矩阵 $R^d_{\theta,m}$ 相乘做空间变换映射成新的向量 x'，旋转角度为 $m\theta_i$，计算公式如下：

$$R^d_{\theta,m}x=\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\\vdots\\x_{d-1}\\x_d\end{pmatrix}\otimes\begin{pmatrix}\cos m\theta_1\\\cos m\theta_1\\\cos m\theta_2\\\cos m\theta_2\\\vdots\\\cos m\theta_{d/2}\\\cos m\theta_{d/2}\end{pmatrix}+\begin{pmatrix}-x_2\\x_1\\-x_4\\x_3\\\vdots\\-x_d\\x_{d-1}\end{pmatrix}\otimes\begin{pmatrix}\sin m\theta_1\\\sin m\theta_1\\\sin m\theta_2\\\sin m\theta_2\\\vdots\\\sin m\theta_{d/2}\\\sin m\theta_{d/2}\end{pmatrix} \quad (2)$$

具体来说，旋转位置嵌入(RoPE)指将 transformer 的词嵌入向量 q、k 旋转到其位置索引的角度倍数上，正是由于这种独特的位置编码技术使 RoFormer 模型可以附带位置间的相对关系，进而更好地解决一词多义引发的易混淆问题。

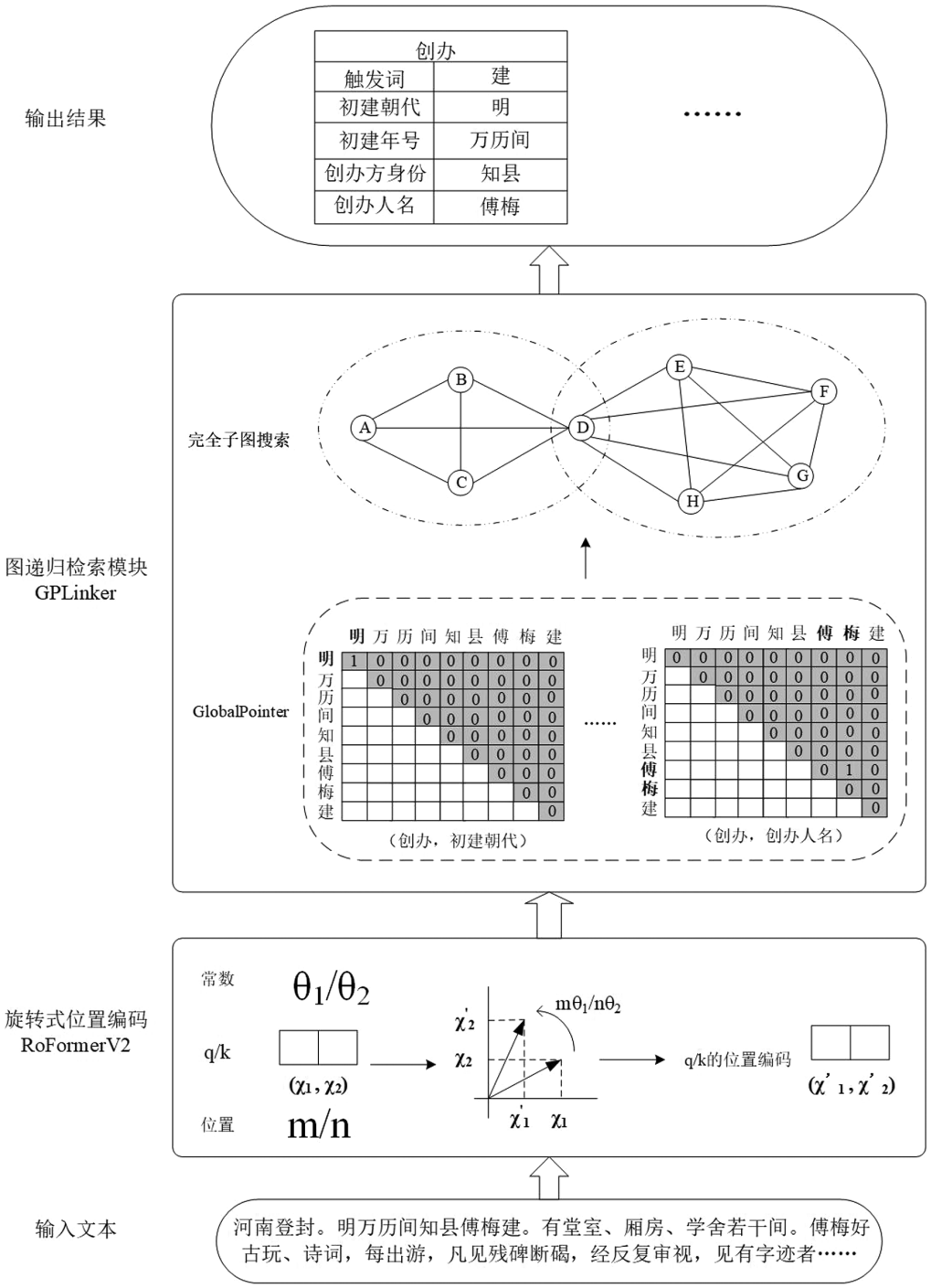

图 1　模型架构图

相比 RoFormer 模型，本文采用的 RoFormerV2 模型做了如下改动[23]：在结构上，RoFormerV2 模型去掉了所有偏置项，并将 transformer 常用的层级归一化调整为均方差层归一化（Root Mean Square Layer Normalization，RMS Norm），删除 RMS Norm 的 gamma 参数；无监督训练方面，RoFormerV2 模型从零开始共使用 280G 数据进行训练，RoFormer 模型则在 RoBERTa 权重基础上，用到的数据仅有 30G 左右；有监督训练方面，RoFormerV2 模型增添了 77 个总计 20G 的标注数据集，进行 92 项如文本分类、指代消解、阅读理解、信息抽取等自然语言处理任务，达到了速度和效果的提升。

2.3 图递归检索模块

图递归检索模块将触发词识别、事件类型分类、论元提取、论元角色分类四个子任务转化为一个联合抽取任务，使触发词作为事件的论元角色，对“(事件类型，论元角色，论元)”的识别转为对“(事件类型，论元角色/触发词，论元/具体触发词)”的组合提取。为有效完成事件抽取任务，本文的图递归检索模块 GPLinker 主要由利用全局归一化的 GlobalPointer 和完全子图搜索两部分构成。

传统的指针网络一般分别识别实体的开始位置和结束位置，造成训练和预测的不一致性，而 GlobalPointer 的设计思想是将实体的开始位置和结束位置作为一个整体处理。假设长度为 n 的文本序列经过 RoFormerV2 模型编码后变为向量序列 $[h_1,h_2,\cdots,h_n]$，对于区域跨度为 $s[i:j]$ 的实体类型 α，将通过公式(3)、(4)两个线性变换得到序列向量的开始位置 $q_{i,\alpha}$ 和结束位置 $k_{i,\alpha}$，其中 α 对应“(事件类型，论元角色/触发词)”这一组合，W_q 和 W_k 为权重矩阵，b_q 和 b_k 是偏置项，实体类型 α 的评分函数如公式(5)所示：

$$q_{i,\alpha}=W_{q,\alpha}h_i+b_{q,\alpha} \tag{3}$$

$$k_{i,\alpha}=W_{k,\alpha}h_i+b_{k,\alpha} \tag{4}$$

$$s_\alpha(i,j)=q_{i,\alpha}^{T}k_{j,\alpha} \tag{5}$$

$s_\alpha(i,j)$ 是用 $q_{i,\alpha}$ 和 $k_{i,\alpha}$ 的内积来表示从第 i 个到第 j 个序列所组成的连续字符串，打分最高的实体将被选中，如图 2 中 GlobalPointer 读取的文本序列为“明万历间知府傅梅建”，$s_\alpha(7,6)$ 表示从 0 开始，纵向第 6 位，横向第 7 位的实体内容“傅梅”，此时对应的实体类型 α 为“(创办，创办人名)”。

针对一个文本序列的“论元/具体触发词”可能属于不同事件的情形，本文使用完全子图搜索方法，将 GlobalPointer 识别得到的(事件类型，论元角色/触发词，论元/具体触发词)作为完全图的节点，而同一事件类型的任意两个节点都可连上边成为相邻节点，组成完全图。如果一个节点同时存在两个子图中，表示该节点的首(事件类型)和尾(论元/具体触发词)都能匹配上，则该节点虽然属于两个事件，但有共同的“事件类型”和“论元/具体触发词”。如香港的周王二公书院，“书院曾于乾隆九年(1744)、道光四年(1824)、1935 年、1965 年 4 次重建”，该句包含四个事件分别为“乾隆九年(1744)重建”“道光四年(1824)重建”“1935 年重建”“1965 年重建”，此时它们存在共同的首和尾，事件类型都为“捐建修缮”，具体触发词也都是“重建”。

图 2 GlobalPointer 结构图

完全子图搜索步骤为：对每一个文本，首先列出所有节点对，如果节点对都相邻，表示该图符合完全图标准，可直接返回。若节点对不相邻，则针对每一对不相邻的节点都分别列出与之相邻的节点集形成子图，再对每一个子图递归检索，从而获得该文本序列涵盖的所有事件论元和具体触发词。

3 实验结果分析

3.1 语料来源与预处理

(1)标注流程

数据集来源于《中国书院辞典》，标注工具为 Excel，标注者仅笔者本人，整体的标注流程包括数据预处理、数据试标注、正式标注、标注结果复验四个阶段，如图 3 所示。

在数据预处理阶段，主要完成对文件的格式转换与数据清洗，包含将 pdf 文件转为 word 文档，检查是否有错别字或文字疏漏，删除空白行，依照省份、书院名称对每个词条进行归类梳理。

在数据试标注阶段，为了能基于文本的不同主题对事件类型进行归纳，笔者先标注 100 条数据，检查事件类型及论元角色前后是否一致、是否已形成一定规则，可供后续标注做参照。若发现当前数据量过少，标注出的事件类型尚无法覆盖全集，将继续增加试标注数据，直至后续标注不会出现新的事件类型，此刻可对当前已标注出的规范进行总结。

在正式标注阶段，将依据试标注总结出的规律对全部文本进行事件类型、触发词、论元角色和具体论元的提取。

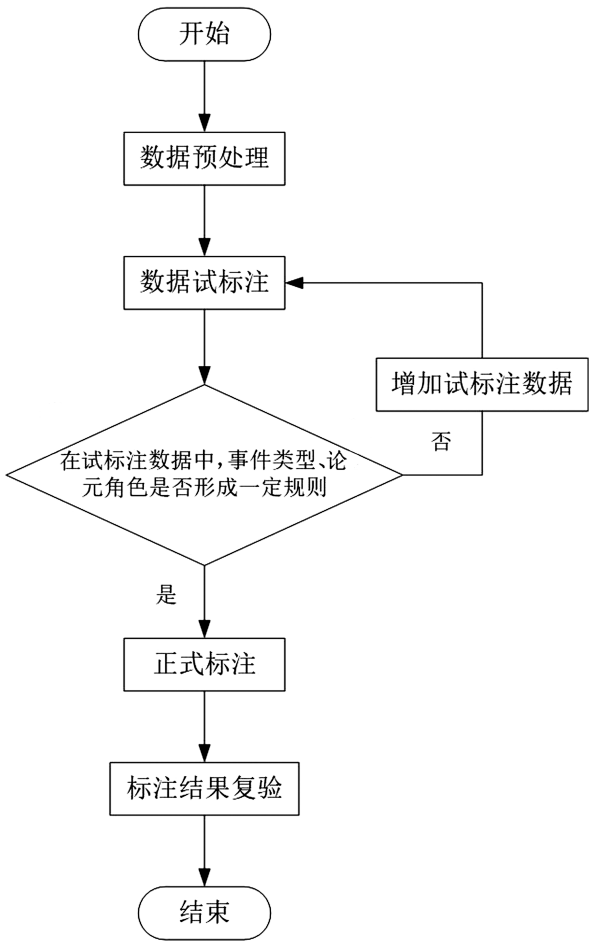

图 3　标注流程图

在标注结果复验阶段，由于目标数据集需包含触发词和论元的起始位，因此用 Excel 内置的 SEARCH 函数查找词语位置，再从第一条开始复查，复查内容包括确定起始位是否准确（同一词条可能含有多个相同的触发词，但 SEARCH 函数只显示第一个字符串匹配的位置），检查事件类型的标注是否遵循规范，核验论元标注是否有遗漏。

（2）事件类型、论元角色的设置

《中国书院辞典》收录了唐代至清代全国各地有史可考的 1626 多所书院资料，对揭示区域性儒家文化的历史传承具有重要价值。该数据集从内容收录角度，既包含动态的叙事性事件，如书院历经朝代更迭所发生的创办、修缮、废除等事件，也涵盖静态的概念性信息，用于描述对象固定不变的状态，如教学管理制度、书院衍生文献。

从动态的叙事性角度来看，书院数据集覆盖创办、捐建修缮、捐田捐资、谕赏赐额、毁坏废除、维护保存、讲学论道等七个事件类型：创办类、捐建修缮类、捐田捐资类事件都相继反映当地官员、宗族势力曾为本县或本族子弟创造求仕条件而建立和修缮书院，为师生提供束脩和膏火，以保证书院的持续运营；谕赏赐额类一般来自于帝王，体现了最高统治者对书院的重视与支持，对书院的嘉奖方式主要包括赐书、赐田、赐额；毁坏废除类与维护保存类分别指对书院造成的破坏和维护，前者多由不可抗因素构成，既有自然灾害一类，如地震、洪水导致书院关闭或被毁，也包含历朝历代的政治动荡加剧了书院的废除，后者则指建国后为保护文化遗存推进的修复和维护工作，其中部分书院已在原址上新建其他机构，本文也将新的机构名记作“书院现名”加以归纳；最后，讲学论道类反映书院内部或不同书院间进行的一种学术交流活动，有宋代朱熹和张栻在岳麓书院持续三天三夜的“朱张会讲”，明代王守仁在贵州文明书院讲学并首次提出知行合一说，这些讲学活动都在一定程度上促进了学派群体的形成和壮大。

从静态的概念性角度而言，书院数据集又有教学内容及管理制度、衍生文献两类：由于其创办目标的差异，教学内容及管理制度也有所不同，教学内容及管理制度类主要覆盖制度变更的具体时间和变更内容，包括书院章程、所教的课程内容；衍生文献类则是以书院为视角产生的一系列文献记录，包括学规、课艺、书院志，反映书院当时的建立、管理和教育理念，这些记录对促进书院领域的研究提供了系统的基础性资料。

在对文本内容总结的基础上，标注形成 9 个事件类型，总计 8740 条事件数据，标注的事件类型、论元角色及数量情况如表 1，事件类型的分布特征如图 4 所示。

由图 4 可知，书院数据集的事件类型在数量上有较大差异，捐建修缮类占比达 51%，其次是创办类，占总事件类型的 19%，毁坏废除类、讲学论道类、捐田捐资分别占 8%、7%、5%，说明《中国书院辞典》整体叙述以时间线为主，主要记载创办、捐建修缮、毁坏废除等对书院发展起到关键作用的事件。相对而言，关于书院的静态描述性信息，如教学内容及管理制度、衍生文献则较少提及，总共仅占 6%。

表 1　事件类型、论元角色的设置

事件类型(数量)	论元角色(数量)
创办(1626)	所在地(1626)、初建朝代(1583)、初建年号(1481)、公元(1238)、创办方身份(1473)、创办人名(1527)、代表性建筑(447)、创建原因(117)、书院别称(173)、书院前身(305)
捐建修缮(4461)	修缮朝代(597)、修缮年号(4211)、公元(2727)、修缮方身份(2999)、修缮方人名(2983)、具体措施(3568)、修缮原因(205)
捐田捐资(453)	捐置年号(448)、公元(241)、捐置方身份(431)、捐置方人名(429)、捐置措施(431)
谕赏赐额(88)	谕赏时间(82)、谕赏内容(87)
毁坏废除(693)	废除时间(691)、废除原因(426)
维护保存(266)	维护时间(20)、书院现名(266)
讲学论道(608)	讲学时间(246)、主讲人(640)、讲学内容(98)
教学内容及管理制度(310)	变动时间(119)、章程(239)、课程(129)
衍生文献(235)	朝代(75)、作者(156)、文献名称(246)

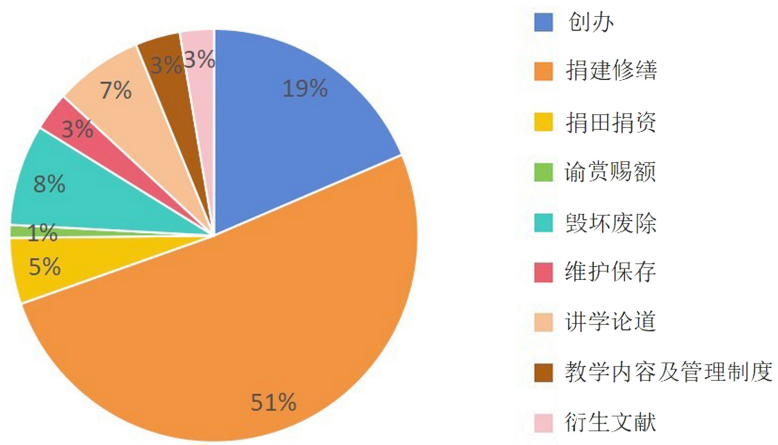

图 4　事件类型的数据分布

(3)缺省补全策略

在标注过程中，笔者发现少量数据存在时间、地点、人名等关键词缺省，书院拟建未成，多个时间论元同时存在的情况。针对上述数据存在的问题，具体处理细则如下：

①时间、地点的缺省补全

时间、地点的上下文引用情况较多，出现缺省的现象也最为普遍。由于汉语篇章上下文连贯度较高，描述事件发生的时间要素、地点要素出现位置相对靠近，时空要素在约束本句的同时有时还会约束后续事件。因此，在时间要素、地点要素的补全策略中，往往使用上一段、上一句的先验时间和地点进行补全，示例如下：

(原文)光绪二十一年(1895)增购新书。三十一年改为房山县高等小学堂。

(补全后)光绪二十一年(1895)增购新书。光绪三十一年改为房山县高等小学堂。

②人名的缺省补全

人名的缺省一般出现在上文已提及人物的具体名称，为避免句子出现重复，提高其表达效率，文章常会出现人物姓或名的缺失，此时可根据上下文进行补全，示例如下：

（原文）清康熙年间，乌石村人进士林琛（任康熙内阁中书）辞官养母，收埋遗骨，重建佛寺和书院于此，集里人讲学其中。佛寺和书院均名“垢洗”，意在洗刷过去污秽。林死后，其子光鼎亦辞官回家，复兴书院，并加以扩展。

（补全后）清康熙年间，乌石村人进士林琛（任康熙内阁中书）辞官养母，收埋遗骨，重建佛寺和书院于此，集里人讲学其中。佛寺和书院均名“垢洗”，意在洗刷过去污秽。林琛死后，其子林光鼎亦辞官回家，复兴书院，并加以扩展。

③不标注无效的事件信息

本文主要记录迁建修缮成功的事件实例，因此书院拟建而未成的内容将不做标注，如“清乾隆间，儒学迁回原址，拟于迎霭门外建横冈书院，未成”。

④多时间要素现象

针对多个时间要素同时存在的情形，本文将只记录事件发生的起始时间，如“清乾隆八年（1743）镇安府知府陈谟购地创兴，乾隆十二年春知府张光宗建成”，以“清乾隆八年（1743）”作为创办类的初建时间。

（4）词条的字数分布情况

标注的词条字数分布情况如表2所示，各书院词条字数由18字至2007字不等，绝大多的书院词条超出128字，多个单句构成的长文本容易造成信息提取缺失。其中94.1%的词条字数小于512字，满足以绝对位置编码的预训练模型的最大输入长度，而剩余5.9%则超出一般预训练模型的训练范围。

表2　词条的字数分布情况

字数区间	书院词条数	词条字数占比（%）
0－128	328	20.17
128－256	787	48.40
256－512	415	25.52
512－1024	87	5.35
1024－2048	9	0.55

3.2　实验参数及评价指标的设置

为有效验证研究方法的有效性，本文选取了近年来以绝对位置编码为主的BERT、RoBERTa、ELECTRA预训练模型，对比RoformerV2模型在同等文本序列识别性能的提升情况，同时在下游设置传统的CRF模型做参照，以验证GPLinker的抽取效果。实验使用的BERT模型为谷歌提供的bert-base-chinese，中文版本的RoBERTa模型与ELECTRA模型为哈尔滨工业大学·讯飞语言认知计算联合实验室发布的RoBERTa-wwm-ext、ELECTRA-base。

实验采用的GPU是NVIDIA Tesla V100 16GB，实验环境为Python 3.7，Tensorflow 2.0，Keras 2.3.1，详细的参数设置如表3所示。

表3　模型参数设置

参数名	参数值
epochs	35
batch_size	8
crf_lr_multiplier	100
learning_rate	1e－5
maxlen	512

文章的事件抽取模型选用准确率（Precision）、召回率（Recall）和F1值（F1－score）作为评价指标，具体计算公式为：

$$P=\frac{\text{识别正确的论元数}}{\text{机器识别的论元数}} \tag{6}$$

$$R=\frac{\text{识别正确的论元数}}{\text{人工标注的论元数}} \tag{7}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{8}$$

3.3　实验结果分析

实验分别设置BERT、RoBERTa、ELECTRA、RoFormerV2四个预训练模型结合CRF、GPLinker模型，按照8:1:1划分训练集、验证集与测试集，单次epoch所耗时间及模型测试结果如表4所示。

在位置编码方面，本文采用旋转式位置编码的实验④、⑧效果均优于使用绝对位置编码的实验组别，实验④相比①、②、③在F1值分别提升了1.49、1.16、2.59个百分点，实验⑧较⑤、⑥、⑦在F1值分别提升了0.96、0.73、1.35个百分点，但单次epoch耗费时间较多。由于attention层本身不具备识别

位置的能力，常规的 BERT、RoBERTa、ELECTRA 模型均显式增加预先定义好的序列位置，而 RoFormerV2 模型用内积形式添加向量的相对位置，使底层 transformer 结构可以同时提取语义和位置信息。实验结果证明，旋转式位置编码的 RoFormerV2 模型具备较好的上下文特征提取能力，更适用于《中国书院辞典》所出现的长文本序列。

表 4　不同模型的效率比较

编号	实验模型	计算时间(s)	P(%)	R(%)	F1(%)
①	BERT－CRF	108	82.78	80.44	81.60
②	RoBERTa－CRF	114	82.49	81.39	81.93
③	ELECTRA－CRF	110	80.76	80.24	80.50
④	RoFormerV2－CRF	318	83.01	83.17	83.09
⑤	BERT－GPLinker	63	89.73	87.25	88.47
⑥	RoBERTa－GPLinker	70	89.98	87.46	88.70
⑦	ELECTRA－GPLinker	64	87.57	88.59	88.08
⑧	RoFormerV2－GPLinker	90	91.15	87.77	89.43

在下游的事件抽取层面，使用图递归检索模块 GPLinker 的实验⑤、⑥、⑦、⑧取得了更好的识别效果，与预训练模型后接 CRF 的实验①、②、③、④相比，在 F1 值上分别提升了 6.87、6.77、7.58、6.34 个百分点，单次 epoch 耗费时间减少，整体计算效率提高。说明 GPLinker 的 GlobalPointer 将实体开始位置和结束位置作为一个整体处理，且后续的完全子图搜索以“首－首”匹配和“尾－尾”匹配模型来构建论元关系，比 CRF 利用特征模板计算联合概率的方法更能识别实体的边界，进而提升模型的性能。

RoFormerV2－GPLinker 模型基于《中国书院辞典》数据集上 35 个 epoch 的 loss 曲线如图 5 所示。该图横轴为训练过程的 epoch，每个 epoch 代表用全部训练数据完整训练了一轮，纵轴为训练的 loss 值。

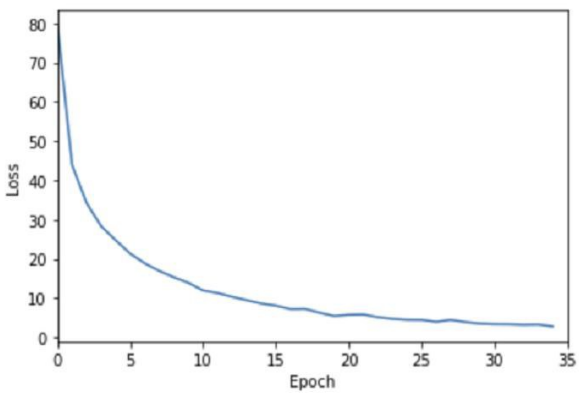

图 5　RoFormerV2-GPLinker 模型的训练 loss 值

从图 5 可以看出 RoFormerV2-GPLinker 模型的 loss 值在前 20 个 epoch 波动比较明显，在后续训练过程逐渐趋于稳定。为了选出最佳模型，在实验中对每个 epoch 产生的训练模型在验证集中加以验证，并记录验证结果，图 6 为模型 35 个 epoch 的训练结果。

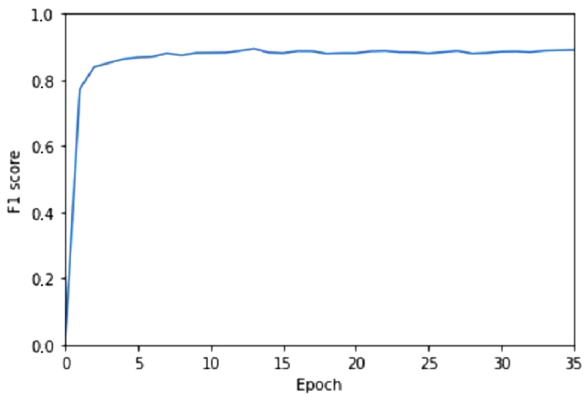

图 6　RoFormerV2－GPLinker 模型的 F1 值

横轴为 epoch，纵轴为验证集的 F1 值，可以看出在第 3 个 epoch 之后，验证集上的 F1 值已逐渐趋于平衡。结合图 4 中的信息，可知在第 20 至 35 个 epoch 之间，模型趋近收敛。

3.4　RoFormerV2－GPLinker 模型的外推性测试

依前文所述，RoFormerV2 模型具备良好的外推性，在理论层面，旋转式位置编码能处理任意长度

的文本序列，而 3.3 节的实验由于受限于 BERT、RoBERTa、ELECTRA 模型的最大文本长度，训练时的参数 max_len 只能设置为 512，无法充分对数据集进行建模。为进一步评估在长文本的外推性能，文章增添了 RoFormerV2－GPLinker 模型于 256、512、1024、2048 四种文本长度的实验，具体结果如图 7 所示。

图 7 利用堆叠直方图的形式绘制不同字数区间在数据集的相对占比情况，用折线图表示模型外推性的效果。由图 7 可知，随着模型所设置的最大文本长度的递增，RoFormerV2－GPLinker 模型能建模的文本数据也在增多，使得 F1 值不断提高。实验结果表明，旋转式位置编码即使在文本长度超过 512 个字符时，识别效果仍能保持提升，该模型通过引入相对位置信息，借助图递归检索方法准确识别实体边界，提取嵌套事件，从而完善对长文本的编码与建模能力，具备较优的外推性。

3.5 案例分析

图 8 展示了 RoFormerV2－GPLinker 模型和 RoBERTa－CRF 模型对“南台书院”抽取的结果分析，所有准确识别的结果都用黄色底纹标出。从图中可以观察到，“光绪二十四年受时务学堂影响，始设立史学、掌故、舆地等课程”的触发词为“设立”，这一触发词大多对应“创办”和“捐建修缮”，而本句其实属于“教学内容及管理制度”类型。由抽取结果可知，RoBERTa－CRF 模型将其分为“捐建修缮”类型，涉及课程变更的部分被列入“具体措施”的事件论元，而 RoFormerV2－GPLinker 模型却能准确识别“设立”和课程内容同属于“教学内容及管理制度”的事件类型，是因为 GPLinker 将触发词和论元列为相邻节点进行递归抽取，降低了误识概率。

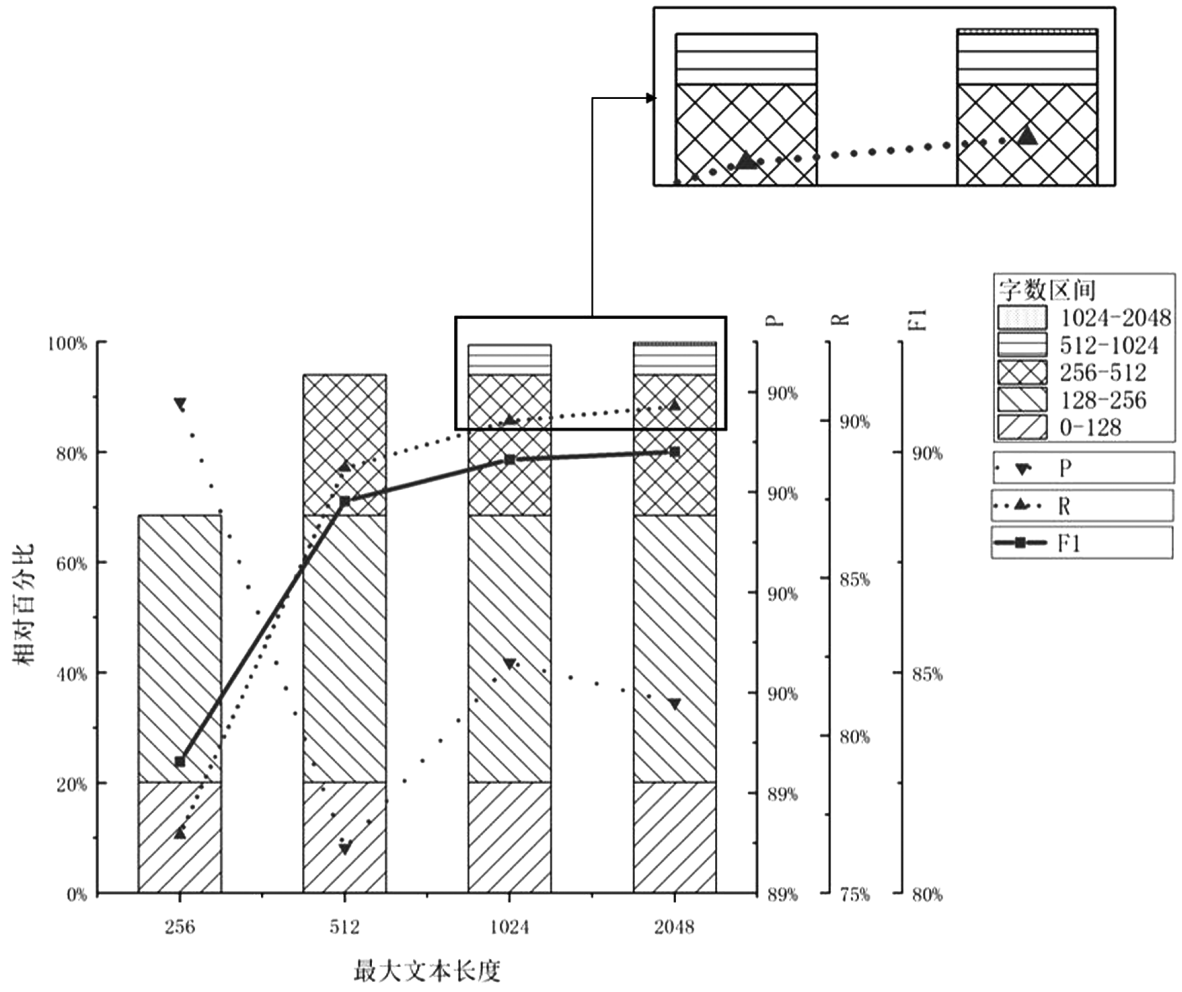

图 7 RoFormerV2－GPLinker 模型的外推性实验

在湖南浏阳。旧为义学在县治横街，元代邑人梁子真建。清康熙三十一年(1692)知县蒋擢迁建西关内。乾隆十年(1745)知县顾维钫重修，改为近圣书院。乾隆二十四年知县张宏燧迁建南台岭，改名‘清浏’。乾隆四十一年始称‘南台’，嘉庆、咸丰、同治间均有修葺，有学田500亩，租800余石。同治六年(1867)增加生童正附课额各2名，时计有生监正附课各12名，童生正附课各22名。光绪二十一年(1895)谭嗣同、唐才常等欲改为算学馆，并‘拟将县中书院改习格致诸学’，未果。光绪二十四年受时务学堂影响，始设立史学、掌故、舆地等课程，时称‘讲舍’，招内课生40名，‘习中学，兼治时务’；外课生80名，‘治西文，必兼中学’。戊戌政变后，一度复旧学。光绪二十八年，改为小学堂。

RoFormerV2-GPLinker模型
事件一：创办
所在地：湖南浏阳
书院前身：义学在县治横街
初建朝代：元代
创办方身份：邑人
创办人名：梁子真
事件二：捐建修缮
修缮朝代：清
修缮年号：康熙三十一年
公元：1692
修缮方身份：知县
修缮方人名：蒋擢
具体措施：迁建西关内
事件三：捐建修缮
修缮年号：乾隆十年
公元：1745
修缮方身份：知县
修缮方人名：顾维钫
具体措施：重修，改为近圣书院
事件四：捐建修缮
修缮年号：乾隆二十四年
修缮方身份：知县
修缮方人名：张宏燧
具体措施：迁建南台岭，改名‘清浏’
事件五：捐建修缮
修缮年号：乾隆四十一年
具体措施：始称‘南台’
事件六：捐建修缮
修缮年号：嘉庆
具体措施：均有修葺，有学田500亩，租800余石
事件七：捐建修缮
修缮年号：咸丰
具体措施：均有修葺，有学田500亩，租800余石
事件八：捐建修缮
修缮年号：同治间
具体措施：均有修葺，有学田500亩，租800余石
事件九：教学内容及管理制度
变动时间：同治六年(1867)
章程：增加生童正附课额各2名，时计有生监正附课各12名，童生正附课各22名
事件十：教学内容及管理制度
变动时间：光绪二十四年
课程：始设立史学、掌故、舆地等课程
事件十一：捐建修缮
修缮年号：光绪二十八年
具体措施：改为小学堂

RoBERTa-CRF模型
事件一：创办
所在地：湖南浏阳
书院前身：义学在
初建朝代：元代
创办方身份：邑人
创办人名：梁子真
事件二：捐建修缮
修缮朝代：清
修缮年号：康熙三十一年
公元：1692
修缮方身份：知县
修缮方人名：蒋擢
具体措施：迁建西关内
事件三：捐建修缮
修缮年号：乾隆十年
公元：1745
修缮方人名：顾维钫
具体措施：重修，改为近圣书院
事件四：捐建修缮
修缮年号：乾隆二十四年
修缮方人名：张宏燧
具体措施：迁建南台岭，改名‘清浏’
事件五：捐建修缮
修缮年号：嘉庆
事件六：捐建修缮
修缮年号：咸丰
事件七：捐建修缮
修缮年号：同治间
事件八：教学内容及管理制度
变动时间：光绪二十四年
事件九：捐建修缮
具体措施：受时务学堂影响
具体措施：始设立史学、掌故、舆地等课程，时称‘讲舍’，招内课生40名，‘习中学，兼治时务’；外课生80名，‘治西文，必兼中学’
事件十：捐建修缮
修缮年号：光绪二十八年
具体措施：改为小学堂

图 8　RoFormerV2－GPLinker 模型和 RoBERTa－CRF 模型的案例研究

4 书院文本事件抽取结果的的可视化分析

4.1 明清书院创办的空间分布特征

在书院发展史上，明代承前启后，清代普及流变，明清时期的书院无论是在数目还是规模都占据主要地位，本文梳理出《中国书院辞典》明清时期“创办”类型所覆盖的地级市名称进而得出书院创办的空间分布统计表。由于该辞典成书于1996年，次年中国设立重庆直辖市和香港特别行政区，1999年又增设澳门行政区，因此文章依据最新行政区划对辞典里的部分省市进行了校正，如表5所示。

表5 明清书院创办的空间分布统计

省/自治区/直辖市	明初（1368－1464年）	明中（1465－1620年）	明末（1620－1644年）	清初（1636－1661年）	清中（1662－1820年）	清末（1821－1911年）
北京市	0	1	1	0	4	2
天津市	0	0	0	0	7	2
河北省	1	12	0	0	31	11
山西省	0	13	0	0	14	8
内蒙古自治区	0	0	0	0	1	2
辽宁省	0	3	0	0	4	7
吉林省	0	0	0	0	1	6
黑龙江省	0	0	0	0	1	3
上海市	0	2	0	0	4	7
江苏省	1	9	0	0	30	14
浙江省	0	17	2	1	26	18
安徽省	0	20	2	1	21	9
福建省	1	6	1	0	19	6
江西省	1	16	1	1	12	9
山东省	2	11	1	0	32	32
河南省	1	25	0	1	45	13
湖北省	0	15	1	1	41	21
湖南省	1	18	0	0	46	32
广东省	2	17	3	1	42	18
广西壮族自治区	1	8	0	0	52	22
海南省	1	9	0	0	10	2
四川省	0	19	1	0	47	15
重庆市	0	2	0	0	12	5
贵州省	0	12	0	0	37	19
云南省	0	20	0	1	40	6
陕西省	0	7	0	2	33	9
甘肃省	0	5	0	0	19	7
青海省	0	0	0	0	4	3
宁夏回族自治区	0	2	0	0	4	4
新疆维吾尔自治区	0	0	0	0	1	1
台湾省	0	0	0	0	19	15
香港特别行政区	0	0	0	0	4	2
澳门特别行政区	0	0	0	0	1	0

据《中国书院辞典》正文所记，明代创办的书院有313所，清代创办的有1010所。由表5可知，明清两代书院的创建均呈现明显的不平衡性，东北、西北地区书院新建数远远少于江南与东南沿海地区。较明代而言，清代书院在原有基础上进一步普及，创办趋势一直呈不断发展壮大之势，尤以河南、安徽、广东、四川、湖南、云南、浙江等省为甚。

明初官学兴盛，朝廷奉行“治国以教化为先，教化以学校为本”的思想，革罢书院，大力倡导官学教育，使书院发展陷入沉寂阶段。明代中期，王、湛之学兴起，自正德至嘉靖年间，湛若水的足迹遍及广东、南直隶（江苏、安徽、上海）等地，王、湛弟子及后生更是于各地开讲会、建书院，致使书院兴建一时大盛。明代后期，国势衰败、政局混乱，朝廷禁毁书院，书院由盛转衰，与明王朝一并走向低谷。

清初顺治时期，为抑制明末民族主义思想，诏令“不许别创书院，群聚徒党”，至康熙年间，朝廷的书院政策逐步放宽，各地纷纷争相兴复书院。雍正、乾隆时期呈现全盛局面，书院步入大力发展期，新建数量众多，涉及区域也更加广泛。道光、咸丰年间，外遭殖民侵略、内历天平天国起义，书院新建的气势渐弱，但仍有发展。之后，洋务运动兴起，西学东渐的思想使书院进入高速普及阶段，覆盖范围较广，以略偏僻的广西地区为例，同治、光绪年间仅书院新建就有15所。

4.2 明清书院建设力量分析

明清两朝官制存在一定差异，明代地方各级行政机构包含省、府、州、县四级，清代实行省、府（州）、县三级行政区划，但在行政级别上又在省和府之间增设道员，清代的州也分为直隶州和散州，直隶州隶属于省，散州则隶属于道或府。文章统计《中国书院辞典》里“创办”类型涵盖的“创办方身份”，从建设力量可分为官方与民间，其中官方力量有中央和地方。清代地方上虽是三级区划，但由于道员名义上与知府平级，行政级别却有隶属关系，而知州也分为直隶州和散州，直隶州知州地位低于知府，散州知州又略高于知县，因此文章中清代的道与州没有列为三级行政区划之一，而是作单独分析，具体可见图9。

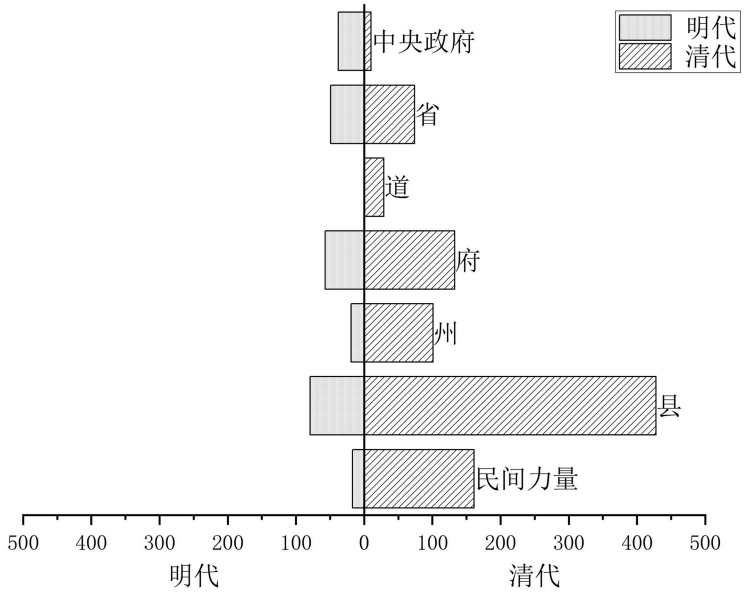

图9 明清书院创办方的身份统计表

明代书院建设主要依赖官方力量，中央政府中的都察院御史、六部尚书、六科给事中等，省级的巡抚、布政使司、按察使司官员，府级的知府、同知，县级的知县、县丞都曾作为书院创办的中坚力量，县级地方官的人数最多，中央、省、府的官员数差异不大，民间力量较少，主要由当地邑人、邑绅组成。在清代书院建设中，依旧依循官办为主、民办为辅的格局，与明代不同的主要有三点：第一，中央级别官员占比在减少，省、道、府、州、县级官员参与变多，说明地方官员在书院创办问题上呈积极态度，使地方官力成为影响书院发展的主要力量；第二，清代知县作为书院创办方的占比在大幅增加，经统计共有422位知

县参与书院兴建，体现县级官员在当地对公共资源的直接管控、分配具备一定的自主权，对书院发展起到重要推动作用；第三，清代的民间力量多了商人和外籍人士的身影，并造就了教会书院这一新的类别，是清代书院的一大特征。

从民间力量的分布来看，创办者的身份大致有以下四种类型，具体如图 10 所示。

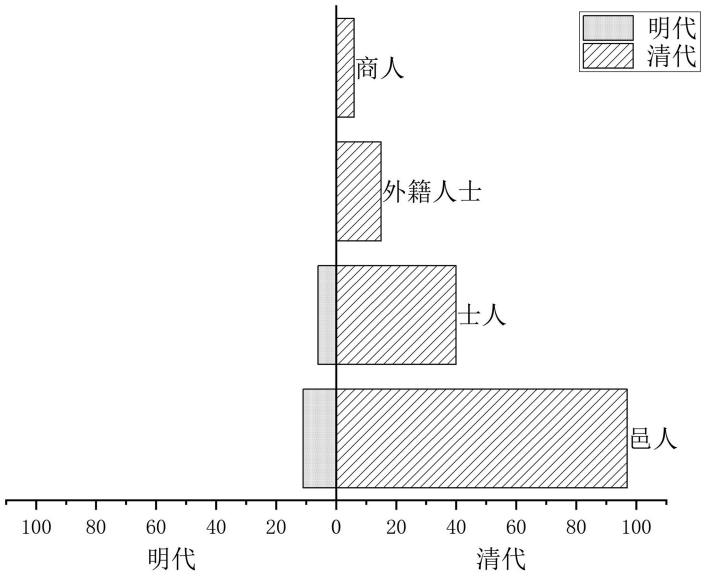

图 10　明清书院创办方的民间身份统计表

明代参与书院创办的民间力量主要由士人与邑人组成，士人多是贡士、进士等科举致仕者，邑人则为当地的儒士、邑绅，他们有的为接续道统、追忆先辈，如状元黄观为纪念南宋爱国志士华岳建翠微精舍，邑人沈国模、史孝咸、管宗圣为祀王守仁创姚江书院，有的为造福乡里、启迪后生，如贡士陈文徽建桐墩书院作为全乡子弟肄业之地，邑人马元吉建明山书院并居院讲薛氏学十余年。

清代书院创办的民间力量仍以士人和邑人为主体，但书院的教学理念却出现了分化，逐步形成博习汉学和经史词章，教授程朱理学，提倡经世致用三类书院，其中以经世致用为指导思想的书院是为结合中西之力，尝试将古老的书院制度和西方近代教育体制接轨，意在推古求新。除士人与邑人群体外，随着商品经济的发展，清代商人的经济实力进一步加强，商人兴资办学的情况屡屡出现，如盐商汪鸣瑞和都转运盐使高熊征集财建紫阳书院，芦商查为义、运使卢见曾建问津书院，都是为子弟舍商而士创造条件。同时，清代书院的创办方也出现了来自美、英、德、法各国的外籍人士，书院所授课程各有不同，有英国长老会牧师莫伟良创立的瞽目书院，专收盲童，教以盲文、算术、音乐等科；美国基督教公理会传教士娄戴德在通县城内建立的八境神学院，课程以圣经、汉文和英文为主；德国基督教同善会传教士尉礼贤创办的礼贤书院，部分课程用德语授课，搜集形象教具并建立化学和物理室。

5　结语

文章综合旋转式位置编码与图递归检索方法构建了书院事件抽取模型，利用 RoFormerV2 模型解决长文本的截断现象，借助 GlobalPointer、完全子图搜索策略化解多论元的嵌套及论元间的关联性问题。研究结果证明，该模型能有效融合向量的位置和语义信息，在事件抽取任务上较基线模型取得了较优结果，并在文本长度超过 512 个字符时，模型识别效果仍能保持提升。进一步地，根据《中国书院辞典》提取出的的事件类型及论元梳理并分析明清书院创办的空间分布及建设力量，发现明清两代书院的创办在区域上均呈现明显的不平衡性，并且都遵循官办为主、民办为辅的格局，但具体在书院覆盖范围和创办人群分布上仍存在较大差异。

文章提出的事件抽取模型融合了向量的位置特

征，以“首－首”匹配和“尾－尾”匹配方式来构建论元关系形成完全图，虽然取得了一定的效果，但未来依然有许多需改进之处。就模型本身而言，可考虑融合更多的词特征信息提升模型性能，去除冗余参数优化模型结构，使模型更轻量化。而在《中国书院辞典》数据方面，未来将辅以更多的古籍文献，对剩余事件类型进行深入分析研究，以挖掘书院的教育活动及其与地方社会的联系。

参考文献

1 季啸风. 中国书院辞典[M]. 杭州：浙江教育出版社，1996.

2 Devlin J，Chang M，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. Minneapolis，USA，2019：4171－4186.

3 Cui Y M，Che W X，Liu T，et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE－ACM Transactions on Audio Speech and Language Processing，2021：3504－3514.

4 Clark K，Luong M T，Le Q V，et al. ELECTRA：pre-training text encoders as discriminators rather than generators[C]//Proceedings of the 8th International Conference on Learning Representation. 2020：1－18.

5 Su J L，Murtadha A，Pan S，et al. Global pointer：novel efficient span-based approach for named entity recognition[EB/OL]. [2024－05－05]. https://arxiv.org/pdf/2208.03054.

6 Chen Y，Xu L，Liu K，et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015：167－176.

7 郭鑫，高彩翔，陈千，等. 面向新冠新闻的三阶段篇章级事件抽取方法[J]. 计算机工程与应用，2023，59(3)：150－157.

8 Wang J，Han B，Wang F，et al. Document-level core events extraction based on QA[J]. Journal of Physics：Conference Series，2022，2171(1)：012062.

9 Li Q，Ji H，Huang L. Joint event extraction via structured prediction with global features[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013：73－82.

10 Nguyen T H，Cho K，Grishman R. Joint event extraction via recurrent neural networks[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. 2016：300－309.

11 葛君伟，乔蒙蒙，方义秋. 基于上下文融合的文档级事件抽取方法[J]. 计算机应用研究，2022，39(1)：48－53.

12 Bethard S，Martin J H. Identification of event mentions and their semantic class[C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing，2006：146－154.

13 Llorens H，Saquete E，Navarro B. TimeML events recognition and classification：learning CRF models with semantic roles[C]//Proceedings of the 23rd International Conference on Computational Linguistics. 2010：725－733.

14 Boros E，Besanon R，Ferret O，et al. Event role extraction using domain-relevant word representations[C]//Proceedings of the Conference on Empirical Methods in Natural Language Proceeding. 2014：1852－1857.

15 Zhang Z，Xu W，Chen Q. Joint event extraction based on skip-window convolutional neural networks[C]//Natural Language Understanding and Intelligent Applications：5th CCF Conference on Natural Language Processing and Chinese Computing. 2016：324－334.

16 Duan S，He R，Zhao W. Exploiting document level information to improve event detection via recurrent neural networks[C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing. 2017：352－361.

17 薛颂东，李永豪，赵红燕. 基于多粒度阅读器和图注意力网络的文档级事件抽取[J]. 计算机应用研究，2024，41(8)：2329－2335.

18 田三川. 基于问答的中文事件抽取研究[D]. 苏州：苏州大学，2022.

19 张虎，张广军. 基于多粒度实体异构图的篇章级事件抽取方法[J]. 计算机科学，2023，50(5)：255－261.

20 余传明，邓斌，谈腊云，等. 基于 XLNET 和 GAT 的句法信息增强事件抽取模型[J/OL]. 数据分析与知识发现：1－18[2024－04－30]. http://kns.cnki.net/kcms/detail/10.1478.G2.20230925.0840.002.html.

21 苏方方，李霏，姬东鸿. 基于可控解码策略的生成式生物医学事件抽取[J]. 中文信息学报，2023，37(11)：68－80.

22 Su J，Lu Y，Pan S，et al. RoFormer：enhanced transformer with rotary position embedding[EB/OL]. [2024－05－05]. https://arxiv.org/pdf/2104.09864v4.

23 苏剑林. RoFormerV2：自然语言理解的极限探索[EB/OL]. [2024－10－11]. https://kexue.fm/archives/8998.

作者贡献说明：
喻雪寒：调研与梳理文献，数据处理与分析，论文撰写
何琳：提出论文选题和总体研究思路，修改论文

作者单位：南京农业大学信息管理系，江苏南京，210095
收稿日期：2024 年 10 月 11 日
修回日期：2024 年 12 月 30 日

（责任编辑：关志英）

Research on the Extraction of Academy Events by Integrating Rotating Position Encoding and Graph Recursive Retrieval Method

YU Xuehan　HE Lin

Abstract: Academies were unique educational institutions in ancient China. The *Chinese Academy Dictionary*, as an important material for recording academies, contained more than 1600 academies that could be examined from the Tang Dynasty to the Qing Dynasty, which was of great value in revealing the historical inheritance of regional Confucian culture. After sorting out the text collection of *the Chinese Academy Dictionary*, we found that this kind of corpus has two characteristics: on the one hand, the entries are based on the academy, and the number of words in some entries exceeds the text input requirements of the conventional pre-training model; On the other hand, there is a phenomenon that different event types share the same trigger word, meaning that one trigger word can represent multiple event types, while the traditional event extraction task regards trigger word recognition as a sequence annotation task, ignoring the correlation between trigger words and event arguments. In order to solve the above problems and comprehensively and systematically sort out and extract the data of academies, based on the review of various modes and methods of event extraction, we developed a comprehensive method integrating rotary position encoding and graph recursive retrieval to extract the event information of academies: the RoFormerV2 model was used to encode the absolute position, so that each vector was attached with the relative position information, and then the event types and arguments were recursively found through the nested entity recognition model GlobalPointer and the complete subgraph search method with the help of the idea of global normalization. Experiments on the *Chinese Academy Dictionary* showed that this method effectively integrated the position and semantic information of vectors and model the relevance between arguments, and overcame the lack of information caused by long texts and the nesting of event arguments, and had good extrapolation ability. Additionally, based on the existing event extraction results, this paper analyzed the spatial distribution and construction strength of the founding of academies in the Ming and Qing Dynasties, and found that the founding of academies in the Ming and Qing Dynasties was obviously imbalanced in regions. Both followed the pattern of being government-owned, but there were still large differences in the coverage of academies and the distribution of founding groups between the two periods.

Keywords: *Chinese Academy Dictionary*; Event Extraction; RoFormerV2; GlobalPointer; Graph Recursive Retrieval