面向数字人文的《四库全书》预训练模型构建及应用研究*

王东波[1]，刘畅[1]， 朱子赫[1]，刘江峰[1]，胡昊天[2]，沈思[3] ，李斌[4]

（1 南京农业大学信息管理学院，江苏，南京，210095；2 南京大学信息管理学院，江苏，南京，210023；3 南京理工大学经济管理学院，江苏，南京，210094; 4 南京师范大学文学院，南京 210097）

摘要：数字人文研究需要大规模语料库和高性能古文自然语言处理工具支持。预训练语言模型已经在英语和现代汉语文本上极大的提升了文本挖掘的精度，目前亟需专门面向古文自动处理领域的预训练模型。我们以校验后的高质量《四库全书》全文语料作为训练集，基于BERT 深度语言模型框架，构建了面向古文智能处理任务的 SikuBERT 和 SikuRoBERTa 预训练语言模型。我们设计了面向《左传》语料的古文自动分词、断句标点、词性标注和命名实体识别 4 个下游任务，分别对我们提出的预训练模型和其他三种用于对比的模型（BERT、RoBERTa、GuwenBERT）进行微调，验证模型性能。经过实验，SikuBERT 和 SikuRoBERTa 模型在全部 4 个下游任务中的表现均超越其他预训练模型，表明我们提出的模型具有较强的古文词法、句法、语境学习能力和泛化能力。基于表现最优的 SikuRoBERTa 预训练模型，我们进一步构建了“SIKU-BERT 典籍智能处理平台”，并提供了典籍自动处理、典籍检索和典籍自动翻译三种在线服务。该平台可以辅助文献学、古汉语文学、历史学等领域学者在不具备数据挖掘与深度学习的专业背景下，以简约可视化的方式对典籍文本进行高效率、多维度、深层次、细粒化的组织处理与分析挖掘。

关键词：数字人文；《四库全书》；预训练模型；深度学习

Construction and Application of Pre-training Model of “Siku Quanshu” Oriented to Digital Humanities

Dongbo Wang[1]，Chang Liu[1]， Zihe Zhu[1]，Jiang Feng[1]，Haotian Hu[2]，Si Shen[3] ，Bin Li[4]

(1 College of Information Management, Nanjing Agricultural University, Nanjing 210095, China ;2 School of Information Management, Nanjing University, Nanjing 210023, China;3School of Economics & Management，Nanjing University of Science and Technology ,Nanjing 210094;4 College of Literature, Nanjing Normal University, Nanjing 210097, China)

Abstract: Digital humanities research needs the support of large-scale corpus and high-performance ancient Chinese natural language processing tools. The pre-training language model has greatly improved the accuracy of text mining in English and modern Chinese texts. At

* 本文系国家自然科学基金面上项目“基于典籍引得的句法级汉英平行语料库构建及人文计算研究”（项目编号：71673143）的研究成果之一.

作者简介：王东波，男，1981 年生，博士，教授，博士生导师，研究领域为自然语言处理与文本挖掘、数字人文下的典籍知识挖掘，信息计量，通信邮箱：db.wang@njau.edu.cn;刘畅，男，1998 年生，硕士研究生，研究领域为自然语言处理与文本挖掘；朱子赫，男，1996 年生，硕士研究生，研究领域为自然语言处理与文本挖掘；刘江峰，男，1998 年生，硕士研究生，研究领域为自然语言处理与文本挖掘；胡昊天，男，1997 年生，博士研究生，研究领域为自然语言处理与文本挖掘；沈思，女，1983 年生，副教授，博士生导师，研究领域为机器学习、信息检索；李斌，1981 年生，副教授，硕士生导师，研究领域计算语言学、典籍知识挖掘。

present, there is an urgent need for a pre-training model specifically for the automatic processing of ancient texts. We used the verified high-quality “Siku Quanshu” full-text corpus as the training set, based on the BERT deep language model architecture, we constructed the SikuBERT and SikuRoBERTa pre-training language models for intelligent processing tasks of ancient Chinese. We designed four downstream tasks: automatic word segmentation, segmentation punctuation, part-of-speech tagging, and named entity recognition for the ZuoZhuan corpus. We performed fine-tuning to verify the performance of the pre-training model we proposed and the other three models for comparison (BERT, RoBERTa, GuwenBERT). After experiments, the performance of SikuBERT and SikuRoBERTa models in all four downstream tasks surpassed other pre-training models, indicating that our proposed model has strong ancient Chinese lexical, syntactic, context learning capabilities and strong generalization capabilities. Based on the best-performing SikuRoBERTa pre-training model, we have further constructed the “SIKU-BERT Classics Intelligent Processing Platform” and provided three online services: automatic processing of classics, retrieval of classics, and automatic translation of classics. The platform can assist scholars in philology, ancient Chinese literature, history and other fields to perform efficient, multi-dimensional, in-depth, and detailed analysis of classic texts in a simple and visual manner without a professional background in data mining and deep learning.
Keywords: Digital Humanities; “Siku Quanshu”; pre-training model; deep learning

1 引言

在人文社会科学的研究中，数字人文研究是近几年发展极为迅速的一个研究方向。在整个数字人文研究当中，对于数字人文概念、研究状况、研究机构的探究相对比较全面和充分，而对于相关语料库、数据库、知识库和计算模型构建并进行发布的研究相对比较少。对于汉语语境下的数字人文探究来说，最大的优势和特点之一是拥有海量的古代典籍数据，这其中比较有代表性的图书典籍数据源为《四库全书》。如何基于《四库全书》这一海量的数据，结合当前深度学习的技术，构建独特的预训练模型对于推进整个古文智能化和数字人文往纵深发展具有独特的意义和价值。在上述这一背景下，结合目前已有的深度学习技术，基于《四库全书》这一海量的图书资源，本文构建了《四库全书》预训练模型，并对模型进行了各个层面上的验证，同时搭建了面向典籍古文处理的应用平台。

2 文献综述

自然语言处理和文本的研究包括序列标注、自动分类、文本生成等各类有监督任务。这些任务往往需要构建大规模标注训练集，以让深度学习模型充分学习词汇、句法与语义特征，人力与时间成本非常昂贵。而通过无监督或自监督的方式，可让语言模型在大量未标记语料上进行训练，对自然语言内在特征进行建模与表征，得到具有通用语言表示[1]的预训练模型（Pre-trained Model, PLM）。在进行下游任务时，直接将预训练模型作为初始化参数，不仅使得模型具备更强的泛化能力与更快的收敛速度[2]，且仅需要输入少量的标记数据进行微调，即可在避免过拟合的同时显著提升 NLP 任务性能。

早期以 Word2Vec[3]、GloVe[4]等为代表的预训练模型基于词嵌入技术，将词汇表征为低维稠密的分布式向量。这些嵌入方式虽然考虑了词义与词汇间共现关

系，但是所构建的词向量为缺乏上下文依赖的静态向量，词义不会因语境的更改而变化，因此无法解决一词多义问题。而自 ELMo[5]模型提出以后，基于上下文语境信息动态嵌入的预训练模型解决了静态词向量词义固定的问题，实现了对词义、语法、语言结构的联合深层建模。

预训练模型根据建模思想的不同，主要可以分为三类。第一类是以 GPT[6]为代表的自回归模型。由于本质上为单向语言模型，虽然在生成式任务中表现优异，但是无法同时学习上下文信息。第二类是以 BERT（Bidirectional Encoder Representation from Transformers）[7]为代表的自编码模型。通过掩码语言模型实现了两个方向信息的同时获取，但也因此导致预训练和微调阶段不匹配的问题。第三类是以 XLNet[8]为代表的排序语言模型。此类模型融合了上述两类模型的优势，通过对输入序列的随机排序使得单向语言模型学习到双向文本表示的同时还保证了两阶段的一致性。

以下是当前较为主流的预训练模型。ELMo（Embedding from Language Models）[5]模型通过两层双向 LSTM 神经网络在大规模语料库上预训练，学习词汇在不同语境下的句法与语义信息，并在下游任务中动态调整多义词的嵌入表示，从而确定多义词在特定上下文中的含义。由于其简单的拼接前后两个方向独立训练的单向语言模型，特征融合能力相对较弱。GPT（Generative Pre-Training）[6]模型将 ELMo 模型中的 LSTM 架构替换为特征提取能力更强的单向 Transformer[9]，从而捕捉更长距离的语境信息。但由于仅使用上文信息预测当前词汇，因此更适合于机器翻译、自动摘要等前向生成式任务。在其后续改进型 GPT2.0[10]与 GPT3.0[11]模型中，采用了更大的 Transformer 结构，基于规模更大、质量更高、类型更广的 WebText、Common Crawl 等数据集预训练了更加通用、泛化能力更强的语言模型，并实现无需微调完全无监督的进行文本生成等下游任务。BERT 模型的出现极大的推动了预训练模型的发展[12]，催生了一系列改进的预训练模型，也使得预训练结合下游任务微调逐渐成为了当前预训练模型的主流模式[1]。BERT 是一种基于 Transformer 架构的自监督深层双向语言表示模型。它通过掩码语言模型迫使模型根据前后文全向信息进行预测，从而实现深层双向文本表示。此外，BERT 还通过下一句预测任务学习前后两个句子是否为连续关系，从而更好的实现自动问答和自然语言推理。

由于 BERT 模型中 MLM 遮蔽机制仅对单个字符进行遮蔽，对词间关系与中文词义的学习并不友好，因此一些预训练模型对遮蔽机制进行了改进。ERNIE（Baidu, Enhanced Representation through Knowledge Integration）[13]在原始对单个字符（汉字）遮蔽的基础上增加了实体层面遮蔽和短语层面遮蔽，从而使预训练模型学习到丰富的外部实体和短语知识。该模型还构建了对话语言模型（Dialogue Language Model, DLM）任务，基于百度贴吧的对话数据学习多轮对话中的隐式语义关系。BERT-wwm[14]模型提出了更适合中文文本的全词遮蔽。不同于 ERNIE（Baidu）仅遮蔽实体和短语，该模型进一步放宽了遮蔽的条件，即只要一个中文词汇中的部分汉字被遮蔽，就把该词汇中的所有汉字全部遮蔽，从而使预训练模型学习到了中文词汇的词义信息。SpanBERT[15]则采用 Span Masking 方法，从几何分布中采样 Span 的长度，并随机选择遮蔽的初始位置，让模型仅根据 Span 的边界词和 Span 中词汇位置信息预测被遮蔽词汇。实验证明该方法表现优于对实体和短语进行遮蔽。RoBERTa（a Robustly Optimized BERT Pretraining Approach）[16]模型将词汇静态遮蔽（static mask）替换成动态遮蔽（dynamic mask），在每次输入前均对句子进行一次随机遮蔽，从而提升了训练

数据的利用率。此外，该模型在预训练过程删去 NSP 任务，改用 FULL-SENTENCES 方法每次输入指定长度的连续句子，进一步优化了模型在句子关系推理方面的表现。StructBERT[17]模型则增加了词汇结构预测（Word Structural Objective）任务，对于输入句中未被遮蔽的词汇，随机选择三个连续的词（Trigram）打乱循序，要求模型重构并恢复先前的顺序。将 NSP 任务替换为句子结构预测，将判断是否为连续句子的二元分类问题改进为预测下一个句子与当前句子位置关系的三元分类任务，从而显式的学习词汇和句子层面的语义关系与语言结构。

部分预训练模型对 BERT 的模型架构进行了修改。为了让结构化的外部知识增强语言表征，ERNIE（THU, Enhanced Language Representation with Informative Entities）[18]模型将知识图谱中的命名实体作为先验知识引入 BERT 的预训练过程。该模型分别采用 T-Encoder 和 K-Encoder 对文本和实体知识进行编码与特征融合，并在预训练过程引入词汇-实体对齐任务，帮助更好的将实体知识注入文本表示中。为了解决 BERT 忽略了被遮蔽词汇间相关性这一问题，XLNet 提出了双流自注意力机制，采用排序语言模型的思想，通过因式分解序列所有可能的排列方式，每个词汇都可学习到两边所有词汇的信息，使得单向的自回归模型也具备了同时学习上下文特征的能力。引入自回归模型 Transformer-XL 中的片段循环机制和相对位置编码，实现对长期依赖关系的学习。由于整个预训练过程并不会将人为遮蔽纳入计算，因此 XLNet 不存在预训练与微调两阶段不匹配的情况。ELECTRA（Efficiently Learning an Encoder that Classifies Token Replacements Accurately）[19]引入了替换标记检测任务，在对输入句进行随机词汇遮蔽后，通过生成器预测词汇并替代标记，随后采用鉴别器分辨生成器产生的词汇是否与原始输入词汇相同，最终仅使用预训练的鉴别器开展下游任务。ELECTRA 解决了预训练任务与下游任务中[MASK]不匹配的问题，在提升计算效率的同时取得更优的表现。DeBERTa（Decoding-enhanced BERT with disentangled attention）[20]模型，提出分解注意力机制，在计算词间注意力权值时，采用解耦矩阵同时考虑词汇间的内容和相对位置信息，融入了词汇间依赖关系。通过增强的掩码解码器嵌入词汇在句子中的绝对位置信息，获得词汇的句法特征。此外，提出虚拟对抗训练算法 SiFT(Scale-invariant-Fine-Tuning)用于提升微调下游任务时模型的泛化能力。与动辄含有上亿个参数的预训练模型相比，ALBERT[21]模型通过嵌入参数矩阵分解以及跨层参数共享的方式显著压缩了参数数量，并将 BERT 中的 NSP 替换为 SOP（Sentence-Order Prediction）任务，用于学习相邻句子间连贯性与衔接关系。

此外，还有一些模型仅部分采用了 BERT 的架构或思想。MT-DNN（Multi-Task Deep Neural Networks）[22]模型是一种用于自然语言理解的预训练模型。它采用多任务学习的思想，在预训练阶段通过共享层基于 BERT 进行词汇与语境嵌入，在微调阶段引入单句分类、文本相似度、配对文本分类和相关性排序等多个任务联合学习，减少模型在特定任务上的过拟合，并更适用于一些缺少标注数据的下游任务。受此启发，基于持续多任务学习的思想，百度在 2020 年发布了预训练模型 ERNIE 2.0 (Baidu)[23]。在保留 BERT 的字符嵌入、句子嵌入和位置嵌入三种嵌入的同时引入任务嵌入，通过增量学习的方式使得模型逐步学习词法、句法、语义层面的 7 种任务，不断提升语言表征能力。T5（Text-To-Text Transfer Transformer）[24]模型基于迁移学习思想，构建了一种文本到文本的 NLP 任务统一框架，使得可以使用相同的模型、损失函数、超参数设置等开展机器翻译、自

动问答、文本分类等任务。

从上述相关研究可以发现，第一，目前大多数预训练模型都是基于大量通用语料训练的，第二，相当一部分预训练模型都是基于 BERT 的改进版本。这些模型虽然普适性强，但是在面向特定领域文本的自然语言处理任务时发挥容易受限。尤其是古代汉语在语法、语义、语用上于现代汉语存在较大差异，即使是面向中文构建的 BERT-wwm 也难以达到在通用语料上的性能水准。此外，虽然已经出现了面向生物医学 BioBERT [25]、临床医学 ClinicalBERT[26]、科学 SciBERT[27]、专利 PatentBERT[28]等特定领域的预训练模型，但是，目前仅有 GuwenBERT①基于继续训练将 BERT 迁移至古汉语，且由于语料规模、简繁转换等因素的限制效果不尽如人意。在古汉语领域，由于缺乏大规模纯净的古文数据，构建古文标注训练集成本高昂，对标注人员具有较高要求。因此，构建高质量无监督古文数据集，训练面向古文自然语言处理任务的预训练模型，对高效开展古文信息处理下游任务研究，拓展数字人文研究内涵，增强社会主义文化自信具有重要意义。

中国拥有卷帙浩繁的古代文献典籍，它们蕴含着中华民族特有的精神价值与文化知识。自 20 世纪 80 年代以来，古籍数字化建设工作举得了不俗的实绩。然而，数字化古籍研究仍面临三重困境：其一，古籍数字化仍囿于整理范畴，更深的知识层次研究尚不充分[29]；其二，现有的古籍利用仍以检索浏览为主，深度利用率较低[30]；其三，国内学界虽占有大量数据，却难以引领古籍的数字研究范式[31]。在数字化时代，古籍研究亟待实现范式革新。源自“人文计算”的数字人文理念与古籍数字化研究之间的深度融合正引起学界的广泛关注。数字人文是“一种代表性实践”，“这种代表性的实践可一分为二，一端是高效的计算，另一端是人文沟通”，其主要范畴是“改变人文知识的发现（Discovering）、标注（Annotating）、比较（Comparing）、引用（Referring）、取样（Sampling）、阐释（illustrating）与呈现（representing）”[32]。数字人文的理论逻辑与技术体系“能够为古籍文献的组织、标引、检索与利用提供新的方法与模式”[29]，“协助学者进行多维度的统计、比较、分析，产生新的知识和思想”[32]，为古籍研究与利用提供新的范式。以《四库全书》数字化为研究对象，本文构建了一种全新的 SIKU-BERT 典籍智能处理平台，重点开发其在典籍自动处理、典籍检索和典籍自动翻译三个方面的功能，在数字人文理念引领下提升深度学习模型对古文语料的准确理解、基于古文语料的 NLP 研究。

3 《四库全书》预训练模型构建

3.1 数据源简介

《四库全书》，又称《钦定四库全书》，是清代乾隆时期编修的大型丛书。基于深度学习技术，本文所使用的《四库全书》为文渊阁版本的。本次实验的训练集共纳入字数 536097588 个，其中去除重复字后共包含汉字 28803 个。数据集内的汉字均为繁体中文。数据集较《四库全书》全文字数少的原因在于本实验去除了原本中的注释部分，而仅纳入正文部分。

四库全书数据集主要有四部分组成——经部、史部、子部、集部。而这四个部分分别由 679、568、897、1262 本书组成。下表 1 展示了四库全书各部各书字

①https://github.com/ethan-yt/guwenbert

数的概况。由此表可见，史部中平均每本书的字数最多，子部和集部次之，而经部最少。从字数分布差异上来看，集部的字数分布差异最小，而史部最大。从单本字数极值来看，子部的单本字数最大值最大，而经部的单本字数最小值最小。

从去重后的字数统计结果来看，集部、史部的用字较多。集部主要包括文学作品，而文学作品的用字往往比较丰富，用词比较凝练，因而其总体字数不多，但去重后用字数反而处于相对高位水平。史部主要包括各类历史著作，这类描述历史人物事件的书籍通常篇幅较大，而由于其中经常出现的人名、地名中会包含一些生僻字，因而其去重后用字依然相对较多。

表 1 四库全书各部下著作字数及不重复字数概况

	部	文献数	均值	标准差	最小值	最大值
字数	经	679	106795	184605.8	231	3147380
	史	568	288400	628212.7	1414	5368247
	子	897	139161	462590.2	270	7918752
	集	1262	138625	304956	1506	4256469
不重复字数	经	679	2804.27	2369.195	137	24299
	史	568	3505.99	1923.194	223	10654
	子	897	2736.85	1903.062	107	13392
	集	1262	3917.51	1594.016	656	12510

3.2 预训练模型构建

3.2.1 预训练模型的构建流程

下图 1 展示了从语料预处理到下游任务验证的全过程。

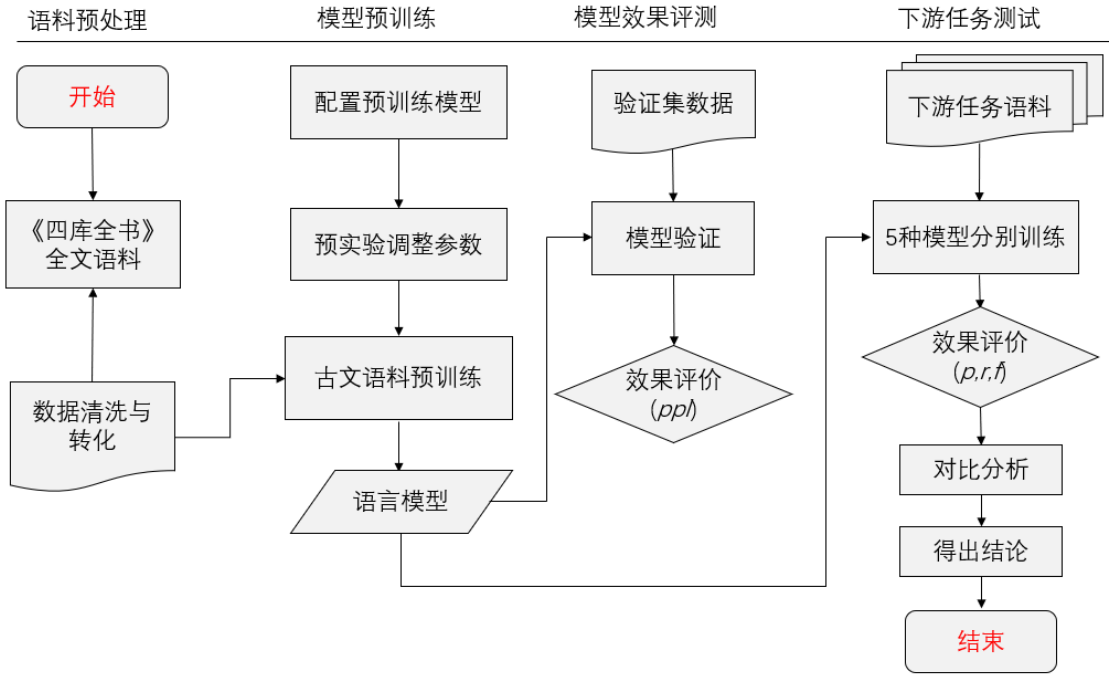

图 1 预训练模型构建实验流程

实验可以分为以下四个部分：语料预处理，语言模型预训练，语言模型效果评测和下游任务测试分析。实验先根据清洗后的《四库全书》全文语料，按 99：1 的比例划分训练集与验证集。模型预训练阶段，在总结多次预实验结果后对训

练参数进行调优，选取 Huggingface 提供的 Pytorch 版 BERT-base-chinese 和 chinese-roBERTa-wwm-ext 模型在训练集上使用掩码语言模型（Masked Language Model）任务完成模型的预训练。在语言模型效果评测阶段，使用困惑度（Perplexity）为基本指标初步判断训练效果，最后通过设置 4 种下游任务进一步分析对比 5 种预训练模型的表现。

3.2.2 预训练模型选取

1）BERT 预训练模型

2018 年 10 月，谷歌 AI 团队公布了一种新的语言表征模型 BERT[7],模型刷新了 11 项 NLP 任务的记录。BERT 的基本结构建立在双向 Transformer 编码器上，通过掩码语言模型（MLM）和下一句预测（NSP）两个无监督任务完成模型的预训练，其中，在 MLM 任务，按比例随机遮蔽输入序列中的部分字符，使模型根据上下文预测被遮蔽的单词，以完成深度双向表征的训练。而在 NSP 任务中，BERT 模型成对地读入句子，并判断给定的两个句子是否相邻，从而获得句子之间的关系。BERT 模型的微调过程则建立在预训练得到的模型上，仅需模型的高层参数进行调整，即可适应不同的下游任务。在本实验中，我们选取了 12 层，768 个隐藏单元，12 个自注意力头，1.1 亿个参数的 BERT 中文预训练模型用于预训练。

2）RoBERTa 中文预训练模型

Liu[16]等人认为 BERT 模型并没有得到充分的训练，为此，该文作者总结了 BERT 模型训练中存在的不足，提出了 RoBERTa 预训练模型。即在 BERT 模型训练的每部分都进行了轻微改进。这些改进包括使用动态掩码替代静态掩码，扩大训练批次与数据集大小，提升输入序列长度以移除 NSP 任务，通过这些调整使 BERT 模型的调参达到了最优。我们选取了 12 层，768 个隐藏单元，12 个自注意力头的 RoBERTa 中文预训练模型作为基础模型，该模型基于全词遮罩（whole word mask）的训练策略在 30G 大小的中文语料上完成了预训练，在全词遮罩中，如果一个词的部分子词被遮罩，则同属该词的其他部分也被遮罩，此方法有助于模型学习中文文本的词汇特征。

3.2.3 语言模型预训练的方法介绍

本实验选用掩码语言模型（MaskLanguageModel,MLM）任务完成 SikuBERT 与 SikuRoBERTa 的预训练。BERT 模型的预训练过程使用了 MLM 和 NSP 两个无监督任务，其中，设计 NSP 任务的目的在于提升对需要推理句间关系下游任务的效果。但在后续的研究中，有学者发现，NSP 任务对 BERT 模型预训练和下游任务性能提升几乎无效。RoBERTa 模型的开发者对 NSP 任务的效果表达了质疑，并通过更改输入句子对的模式设计了四组实验证实了该猜想。Lan[21]等人认为 NSP 任务的设计过于简单，即将主题预测与相干性预测合并在同一个任务中，主题预测功能使 NSP 的损失函数与 MLM 的损失函数发生了大量重叠。

基于上述研究的结果，本实验移除了 BERT 预训练中的 NSP 任务，仅使 MLM 任务完成 SikuBERT 与 SikuRoBERTa 的预训练。在实验中随机遮罩 15%的词汇，通过预测被遮罩字符的方式完成参数更新，并使用 MLM 损失函数判断模型预训练的完成度。全部实验均依靠 Transformers 框架进行。

3.2.4 预训练模型效果的评价指标

在模型效果评测阶段，我们使用困惑度（PPL，perplexity）来衡量语言模型

的优劣，困惑度的定义如下：

对于一个给定的序列 S：$S = w_1 w_2 \ldots w_{n-1} w_n$，$w_n$表示序列中第 n 个词，则该序列的似然概率定义为：

$$\mathrm{p(L)} = \mathrm{P}(w_1 w_{2\ldots} w_{n-1} w_n) \quad \text{公式 1}$$

则困惑度可以定义为：

$$\mathrm{PPL} = \mathrm{P}(w_1 w_{2\ldots} w_{n-1} w_n)^{\frac{-1}{n}} \text{公式 2}$$

困惑度的大小反应了语言模型的好坏，一般情况下，困惑度越低，代表语言模型效果越好。在本实验中我们通过调整训练轮次，SikuRoBERTa 在验证集上的困惑度达到 1.4，SikuBERT 的困惑度达到 16.787，初步验证表明，经过领域化语料上的二次微调，SikuBERT 和 SikuRoBERTa 具有较低的困惑度。从评价语言模型的角度来看，在《四库全书》语料下，相比原始 BERT 模型和 RoBERTa 模型其性能有所提升，可以保证模型充分学习到《四库全书》的语言信息。

3.3 预训练模型性能验证

为进一步验证 SikuBERT 和 SikuRoBERTa 预训练模型的性能，我们设置了以下四项 NLP 任务做进一步的验证，分别为：古文命名实体识别任务、古文词性识别任务、古文分词任务、古文自动断句和标点任务。其次，在语料的选择上，我们基于经过人工校对过的《左传》语料，构造了四种实验所需要的训练和测试数据。第三，在基线模型的选择上，除 BERT 和 RoBERTa 外，同时还引入 GuwenBERT 预训练模型进行验证，下面分别对性能验证实验的语料、任务、模型和结果做进一步的说明。

3.3.1 验证实验的语料和任务介绍

验证实验所使用的语料为李斌等人校对过的繁体《左传》，全文 18 万字。基于南京师范大学制定的古汉语分词与词性标注规范和自动分析工具，人工校对了分词和词性信息的先秦《左传》语料库形成了《左传》数字人文数据库的构建[33]。

《左传》数字人文数据库语料过处理后，除词性识别任务外还可用于古文分词、古文实体识别、古文断句和古文标点任务，选用该语料作为验证实验数据，一方面统一了选用语料的来源，避免了多种古文语料间差异带来的验证上的误差，另一方面能够其经过高质量的人工校对，大大降低语料引入的误差，能够更好地比对预训练模型之间的差异。基于《左传》语料的四个任务名及其内容如表 2 所示。

表 2 下游任务语料描述

序号	任务名称	任务内容
1	古文词性标注	识别古文词语的词性
2	古文分词	对古文进行自动分词
3	古文命名实体识别	识别人名、地名、时间实体
4	古文断句与标点	对古文进行自动断句和标点

接下来对这四项任务的内容和数据处理方式进行具体的说明：

1）古文词性标注任务。古籍文本中没有词界，进行词语的切分，能以词为

粒度进行更多的古文应用[33]，例如古文词典编撰、古文检索等。在训练数据预处理上，因为《左传》数字人文数据库本身即是经过人过校对过词性标签的语料集，所以可以直接作为古文词性标注的训练数据进行使用。

2）古文分词任务。《左传》数字人文数据库是以词为单位进行的词性标注，因此在经过词性标签的清洗后，我们获取了古文的分词数据，该分词数据看作是是词性标注数据的子集，同样可以用于序列标注任务。

3）古文命名实体识别任务。古文实体识别能够展示古文中更细的粒度，有助于古文知识图谱、古文检索等应用。在训练数据预处理上，首先，基于词性标注中的名词词性进行筛选，其次，选取其中人名实体、地点实体和时间词实体作为实体识别任务的识别目标，获得了古文实体识别任务的训练和测试数据。

4）古文自动断句和标点任务。首先，在《左传》数字人文数据库的语料的基础上，去除分词和词性标签，保留标点符号，其次，将每个标点符号作为标记，构造断句和标点训练语料，以希望模型能够为原始古文语料进行断句和标点的操作。

3.3.2 验证模型介绍

验证试验选用的预训练模型如表 3 所示。为验证 SikuBERT 和 SikuRoBERTa 的性能，实验选用基线模型为 BERT-base-chinese 预训练模型[②]和 chinese-roBERTa-wwm-ext 预训练模型[③]，同时还引入 GuwenBERT 预训练模型进行验证，GuwenBERT 基于“殆知阁古代文献语料”在中文 BERT-wwm 预训练模型上进行训练，将所有繁体字均经过简体转换处理后用于训练[④]，模型在古文数据的任务中具有良好的表现。此外，为使验证结果具有一致性，在四项任务的验证中，我们只对上游预训练模型进行更换，对下游任务的模型的参数保持统一。

表 3 验证试验选用的预训练模型一览表

序号	预训练模型名
1	BERT-base-chinese
2	GuwenBERT-base
3	SikuBERT
4	chinese-roBERTa-wwm-ext
5	SikuRoBERTa

3.3.3 模型验证性能指标

结合国内外对分词性能评价的常用指标体系，本项目对 BERT、RoBERTa、GuwenBERT、SikuBERT 和 SikuRoBERTa 预训练模型使用以下 3 个指标来衡量，即准确率 P（Precision）、召回率 R（Recall）、F 值（F-measure），各指标具体计算如下。

$$\text{准确率P} = \frac{A}{A+B} \times 100\% \quad \text{公式 3}$$

$$\text{召回率R} = \frac{A}{A+C} \times 100\% \quad \text{公式 4}$$

② https://huggingface.co/bert-base-chinese
③ https://huggingface.co/hfl/chinese-roberta-wwm-ext
④https://github.com/Ethan-yt/guwenbert

$$调和平均值F = \frac{2\times P\times R}{P+R} \times 100\% 公式\ 5$$

本文选用繁体中文版的《四库全书》全文数据进行预训练模型的领域学习实验，并将其应用在语源为繁体中文的《左传》语料上进行古文自动分词实验。

3.3.4 基于预训练模型的分词结果比较分析

表 4 模型分词结果指标平均值

模型	精确率（P）	召回率（R）	调和平均值（F）
BERT-base-chinese	86.99%	88.15%	87.56%
GuwenBERT-base	46.11%	57.04%	50.86%
RoBERTa	80.90%	84.77%	82.79%
SikuBERT	88.62%	89.08%	88.84%
SikuRoBERTa	88.48%	89.03%	**88.88%**

表 4 为模型指标的平均值，从这些数据我们可以看到实验中我们的 SikuBERT 和 SikuRoBERTa 的性能表现最佳，分词的准确率、召回率和调和平均值均较基线模型 BERT、RoBERTa 和 GuwenBERT 有明显改进。针对准确率和召回率，SikuBERT 预训练模型的表现效果最佳，分别为 88.62%和 89.08%；而 SikuRoBERTa 得到了最好的调和平均值，为 88.88%。所有模型中 GuwenBERT 预训练模型的分词表现最差，精确率、召回率和调和平均值分别为 46.11%、57.04%、50.86%。以调和平均值为基准，在分词任务中原始 BERT 模型表现优于 RoBERTa 模型，识别效果约高出 5%，SikuBERT 预训练模型效果最优。

3.3.5 基于预训练模型的词性标注结果比较分析

表 5 模型词性识别结果指标平均值

模型	精确率（P）	召回率（R）	调和平均值（F）
BERT-base-chinese	89.51%	90.10%	89.73%
GuwenBERT-base	73.31%	77.49%	74.82%
RoBERTa	86.70%	88.45%	87.50%
SikuBERT	89.89%	90.41%	**90.10%**
SikuRoBERTa	89.74%	90.49%	90.06%

基于预训练模型的词性标注实验所用数据集来自《左传》，以领域内较常使用的准确率 P（Precision）、召回率 R（Recall）和 F 值（F-measure）三个指标作为最终实验结果评价方式。从实验结果中可以看出，针对《左传》数据的古文词性标注实验结果均表现不错，但 SikuBERT 和 SikuRoBERTa 模型的调和平均值要明显高于其他三个识别模型，二者的 F 值均超过了 90%，SikuBERT 识别效果更是达到了 90.10%。其中 GuwenBERT 模型的识别效果最差，调和平均值只有 74.82%，不及基础的 BERT 模型。实验结果还表明，原始 BERT 模型效果要优于 RoBERTa 模型，且基于《四库全书》数据训练得到的 SikuBERT 模型效果同样优于 SikuRoBERTa 模型，这一实验结果值得进一步分析和探讨。

3.3.6 基于预训练模型的断句结果比较分析

表 6 模型断句识别结果指标平均值

模型	精确率（P）	召回率（R）	调和平均值（F）

BERT-base-chinese	78.77%	78.63 %	78.70%
GuwenBERT-base	46.35%	20.71%	28.32%
RoBERTa	66.71%	66.38%	66.54%
SikuBERT	87.38%	87.68%	**87.53%**
SikuRoBERTa	86.81%	87.02%	86.91%

为验证 SikuBERT 和 SikuRoBERTa 预训练模型对于古文断句的识别效果，我们在《左氏春秋传》《春秋公羊传》和《春秋谷梁传》三本古文著作数据集中进行断句识别实验，实验结果显示 SikuBERT 和 SikuRoBERTa 模型效果均超过的 85%，SikuBERT 的最优 F 值最高达到了 87.53%，同时这也是多组对比实验中的最好实验结果。GuwenBERT 模型的识别调和平均值在各组实验中表现最差，仅有 28.32%，远低于其他识别模型的识别效果。基础的 BERT 和基于原始 BERT 模型训练得到的 RoBERTa 识别效果一般，其调和平均值分别只有 78.70%和 66.54%，低于我们自主预训练的识别模型，但要高于 GuwenBERT 的识别结果。

3. 3. 7 基于预训练模型的实体识别结果比较分析

表 7 模型实体识别结果指标平均值

预训练模型	实体类别	精确率（P）	召回率（R）	调和平均值（F）
BERT-base-chinese	nr(人名)	86.66%	87.35%	87.00%
	ns(地名)	83.99%	87.00%	85.47%
	t(时间)	96.96%	95.15%	96.05%
	avg/prf	86.99%	88.15%	87.56%
GuwenBERT-base	nr(人名)	39.91%	54.10%	45.93%
	ns(地名)	42.36%	50.40%	46.03%
	t(时间)	85.71%	89.55%	87.59%
	avg/prf	46.11%	57.04%	50.86%
RoBERTa	nr(人名)	79.88%	83.69%	81.74%
	ns(地名)	78.86%	84.08%	81.39%
	t(时间)	91.45%	91.79%	91.62%
	avg/prf	80.90%	84.77%	82.79%
SikuBERT	nr(人名)	88.65%	88.23%	88.44%
	ns(地名)	85.48%	88.20%	86.81%
	t(时间)	97.34%	95.52%	96.42%
	avg/prf	88.62%	89.08%	88.84%
SikuRoBERTa	nr(人名)	87.74%	88.23%	87.98%
	ns(地名)	86.55%	88.73%	87.62%
	t(时间)	97.35%	95.90%	96.62%
	avg/prf	88.48%	89.30%	**88.88%**

对于长文本中实体的有效识别，是判断该模型能否有效解决自然语言理解问题的重要评价标注之一。本组对比实验的数据来自《左传》典籍数据，识别实体对象为数据集中的“人名”、“地名”、“时间”实体，模型识别效果评价标准为最常用的 PRF 值。从实验结果中可以看出，SikuBERT 和 SikuRoBERTa 模型的三种实体识别效果均高于其他三种模型，尤其是对于时间实体的识别实验中，SikuBERT 和 SikuRoBERTa 模型识别结果的调和平均值均超过了 96%。而

GuwenBERT 模型在三类实体识别实验中的表现均最差，其中人名和地名实体的识别效果均低于 50%，同时远低于其他组的识别效果。原始 BERT 模型和 RoBERTa 在三组实验中的表现较为中庸，没有展示特别突出的识别性能。

4 基于预训练模型的典籍智能处理平台搭建

4.1 典籍智能处理平台构建流程

本项目在实现基于 BERT、RoBERTa、GuwenBERT、SikuBERT 和 SikuRoBERTa 预训练模型，分别在六种不同的古文任务中进行性能验证后，结果表明 SikuBERT 和 SikuRoBERTa 预训练模型能够有效提升繁体中文语料处理的效果。为了利于古文 NLP 研究，方便文献学、历史学等学科相关研究人员的工作，我们构建了 SIKU-BERT 典籍智能处理平台，其具体平台构建流程如下图所示。

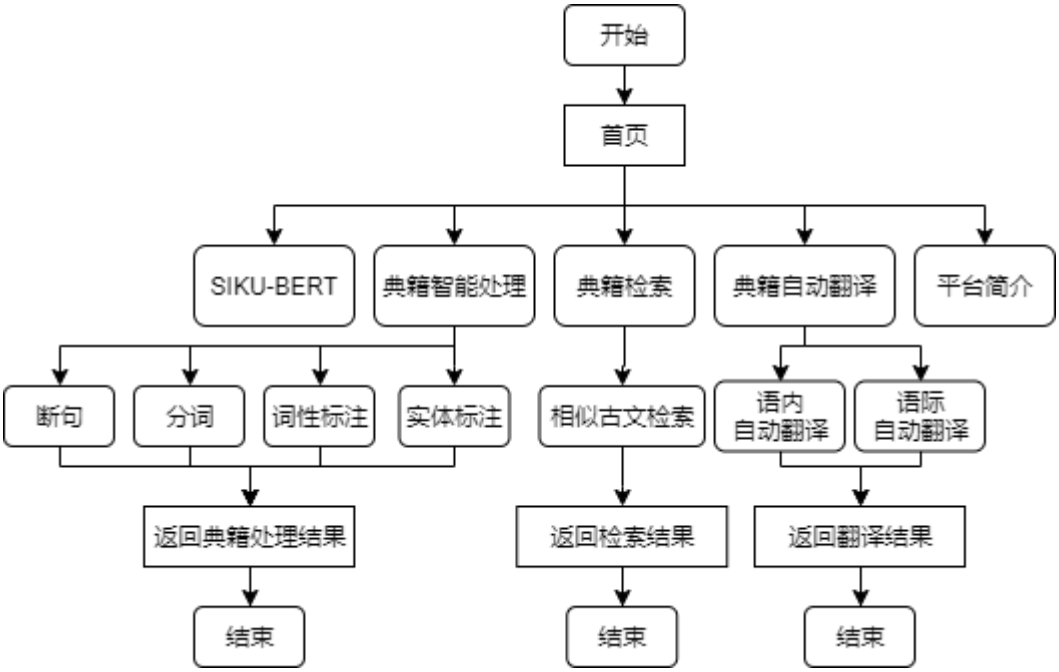

图 2 SIKU-BERT 典籍智能处理平台系统构建流程

如上图 2 所示，本平台共有三种主要功能，包括典籍智能处理功能、典籍检索功能和典籍自动翻译功能。通过首页，用户可以详细了解 SIKU-BERT 的背景，典籍智能处理、典籍检索和典籍自动翻译这三种主要功能的简介以及我们智能处理平台的基本信息。用户可以根据自身需求来选择不同的功能，通过各自的入口进入平台的相应界面。例如，用户希望使用典籍智能处理功能，则可以点击相应界面，选择对输入的典籍文本进行处理的操作（文本断句、分词、词性标注或实体标注），在结果返回框中获得期望的典籍处理结果；针对典籍自动翻译功能，用户可以选择语内翻译或语际翻译，平台将根据用户选择返回翻译结果；如果是选择古文相似检索功能，那么平台将查询的古文句子与语料库中的候选句的相似度进行计算，通过计算相似度的方式来将与查询古文句子相似的古文句子返回。

4.2 平台实现方法和工具

4.2.1 系统数据和工具

考虑到典籍智能处理等功能的实用性，我们搭建了具体的应用平台。在构建方式的选择上，考虑到平台构建的方便性和用户的使用简便性。本平台采用构建一个网站来实现平台。在编程语言上，我们选择了 Python 语言为主要的编程语言。考虑到网站的性能与质量，我们所搭建网站的框架为 Django 框架。该框架是利用 Python 开发的一个免费开源的 Web 框架，几乎包含了 Web 应用的各个方面。在多层级的平行语料库存储上，考虑到存储的资源我们选择 SQLite 数据库进行存储，SQLite 占用的存储资源非常少。最后在本网站前端上，我们选择常用的 HTML、CSS、JS 作为前端的构建工具。

4.2.2 系统功能设计

为了将 SIKU-BERT 模型得到充分的应用，考虑到用户可以快速高效的获取古文断句、分词、词性标注和实体标注的处理结果，相似古文的检索结果以及段落、句子、词汇的语内和语际英文翻译。我们的 SIKU-BERT 典籍智能处理平台主要有典籍自动处理、典籍检索和典籍自动翻译三大功能。

4.2.3 系统应用展示

SIKU-BERT 典籍智能处理平台包含主界面、典籍智能处理界面、典籍检索界面和典籍自动翻译四个界面。主界面通过一目了然的导航栏显示，集成了三大核心功能和平台简介，功能分别包括：典籍智能处理、典籍检索和典籍自动翻译。

1）网站首页

如图所示，打开 SIKU-BERT 典籍智能处理平台网站，可以通过导航栏了解平台的基本信息和首页内容，分别为“Siku-BERT”、“典籍智能处理”、“典籍检索”、“典籍自动翻译”和“平台简介”，对应下方的页面详细介绍。

图 3 SIKU-BERT 典籍智能处理平台网站首页

以“典籍智能处理”为例，本项目主要实现的智能处理功能包括典籍的自动断句、分词、词性标注和实体标注。首页有该功能的运行示例图和详细介绍，通过点击“FIND OUT MORE”进入功能界面。

图 4 SIKU-BERT 典籍智能处理平台首页“典籍智能处理”功能介绍

网站首页底端对本平台的主要功能进行简介，见图 5。

图 5 SIKU-BERT 典籍智能处理平台首页“平台简介”

2）典籍智能处理功能界面

如下图 6 所示，在 SIKU-BERT 典籍智能处理平台的“典籍智能处理”功能界面，用户可以根据自身需求通过上方按钮分别选择对应的典籍处理功能，包括断句、分词、词性标注和实体标注。用户在界面左侧的文本框中输入需要进行处理的原始典籍文本，在选择功能按钮后，可以通过点击“开始处理”的按钮，即可返回经平台处理后的句子。如在上图中输入“子墨子曰：“今若有能以义名立于天下，以德求诸侯者，天下之服可立而待也。’”，在选择“词性标注”功能后，点击“开始处理”，那么在右侧便会输出返回的结果：“子墨子/nr 曰/v：/w“/w 今/t 若/c 有/v 能/v 以/p 义/v 名/n 立/v 于/p 天下/n，/w 以/p 德/n 求/v 诸侯/nr 者/r，/w 天下/n 之/u 服/n 可/v 立/v 而/c 待/v 也/y。/w”。该功能可以让用户可以快速地获得经过规范处理的典籍文本，作为古文 NLP 研究工具，极大地方便了文献学、历史学等学科相关研究人员的工作。

图 6 SIKU-BERT 典籍智能处理平台“典籍智能处理”功能界面

5 结语

在基于古文语料的 NLP 任务中，考虑到异体字现象和“一简对多繁”现象的存在，采用简繁转换功能必然或多或少的丢失古籍中原本的语义信息，因此，使用繁体中文的原始语料仍然是古文自然语言处理的主流，但是，随着预训练模型技术的兴起，面向古文语料的预训练模型并没有得到充分的开发，因此，训练一种能够贴合古文语料的预训练模型一方面能够提升深度学习模型对古文语料的理解，另一方面可以为基于古文语料的 NLP 研究提供支撑，具有重要意义。

本文基于 BERT、RoBERTa、GuwenBERT、SikuBERT 和 SikuRoBERTa 预训练模型，分别在六种不同的古文任务中进行性能验证，实验结果表明，第一，SikuBERT 与 SikuRoBERTa 相较于原本的预训练语言模型的识别效果有一定程度上的提升，SikuRoBERTa 的性能最好。第二，SikuRoBERTa、SikuBERT 在分词，词性标注上的提升幅度较小，在断句，实体识别等任务中的提升幅度较大。

综上，SikuBERT 和 SikuRoBERTa 预训练模型能够有效提升繁体中文语料处理的效果，有利于古文 NLP 研究，方便文献学、历史学等学科相关研究人员的工作。下一步的研究会着手构建更合适古文任务的预训练模型词表，从而获性能更好的词表示特征。

参考文献：

[1]王乃钰,叶育鑫,刘露,凤丽洲,包铁,彭涛.基于深度学习的语言模型研究进展[J].软件学报,2021,32(04):1082-1115.

[2] Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China Technological Sciences, 2020: 1-26.

[3] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[4] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language

processing (EMNLP). 2014: 1532-1543.
[5] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018..
[6] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[EB/OL].[2021-04-24].
https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.
[7] 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.
[8] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. arXiv preprint arXiv:1906.08237, 2019.
[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.
[10] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
[11] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[J]. arXiv preprint arXiv:2005.14165, 2020.
[12]余同瑞,金冉,韩晓臻,李家辉,郁婷.自然语言处理预训练模型的研究综述[J].计算机工程与应用,2020,56(23):12-22.
[13] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.
[14] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinesebert[J]. arXiv preprint arXiv:1906.08101, 2019.
[15] Joshi M, Chen D, Liu Y, et al. Spanbert: Improving pre-training by representing and predicting spans[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 64-77.
[16] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
[17] Wang W, Bi B, Yan M, et al. Structbert: Incorporating language structures into pre-training for deep language understanding[J]. arXiv preprint arXiv:1908.04577, 2019.
[18] Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced language representation with informative entities[J]. arXiv preprint arXiv:1905.07129, 2019.
[19] Clark K, Luong M T, Le Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators[J]. arXiv preprint arXiv:2003.10555, 2020.
[20] He P, Liu X, Gao J, et al. Deberta: Decoding-enhanced bert with disentangled attention[J]. arXiv preprint arXiv:2006.03654, 2020.
[21] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.
[22] Liu X, He P, Chen W, et al. Multi-task deep neural networks for natural language understanding[J]. arXiv preprint arXiv:1901.11504, 2019.
[23] Sun Y, Wang S, Li Y, et al. Ernie 2.0: A continual pre-training framework for language understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 8968-8975.
[24] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. arXiv preprint arXiv:1910.10683, 2019.
[25] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation

model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.
[26] Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission[J]. arXiv preprint arXiv:1904.05342, 2019.
[27] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text[J]. arXiv preprint arXiv:1903.10676, 2019.
[28] Lee J S, Hsiang J. Patentbert: Patent classification with fine-tuning a pre-trained bert model[J]. arXiv preprint arXiv:1906.02124, 2019.
[32] Unsworth J. What Is Humanities Computing and What Is Not?[EB/OL].[2012-09-05]. http: //computerphilologie. tu-darmstadt. de /zeitschriften /framezs. html
[29]李明杰、张纤柯、陈梦石. 古籍数字化研究进展述评(2009-2019)[J]. 图书情报工作，2020，64(06):130-137.
[30]欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报，2016，42(02):66-80.
[31]史睿. 数字人文忧思录[J]. 数字人文，2020(02):157-160.
[32]郑永晓，段海蓉. 古籍数字化、数字人文与古代文学研究——访中国社会科学院郑永晓教授[J]. 吉首大学学报(社会科学版)，2020，41(02):144-151.
[33]李斌,王璐,陈小荷,王东波.数字人文视域下的古文献文本标注与可视化研究——以《左传》知识库为例[J]. 大学图书馆学报,2020,38(05):72-80+90.