基于深度学习的《方志物产》用途实体自动识别模型构建与应用

李娜
（南京林业大学人文社会科学学院，南京 210037）

摘要：以特色馆藏文献《方志物产》为研究语料，基于人工标注语料，运用Bi-LSTM、Bi-LSTM-CRF、BERT、Siku-BERT等4种深度学习模型开展实验，以精确率*P*、召回率*R*、调和平均数*F*作为测试指标，对模型的识别性能进行对比分析，促进物产知识的挖掘和利用。实验结果显示：相较于基于CRF的模型，4种深度学习模型的整体性能取得明显提升；Bi-LSTM、Bi-LSTM-CRF、BERT、Siku-BERT的最好*R*值分别为74.80%、78.05%、88.62%、89.74%；BERT、Siku-BERT注意力机制类深度学习模型的识别效果优于Bi-LSTM、Bi-LSTM-CRF循环类深度学习模型。由于方志类古籍文本结构复杂多样、人工标注精度存在误差、语料规模较小等因素，自动识别模型的实体抽取性能仍有较大的优化空间，但深度学习模型在方志类古籍的内容挖掘中表现出一定的优越性，且不同语料间预训练模型的迁移应用具有可行性。

关键词：深度学习；方志物产；命名实体识别；数字人文；用途实体
中图分类号：K29；N99　DOI：10.3772/j.issn.1673-2286.2022.12.003
引文格式：李娜. 基于深度学习的《方志物产》用途实体自动识别模型构建与应用[J]. 数字图书馆论坛，2022(12)：19-28.

数字人文自诞生以来，便在技术研发与人文研究之间架起了一座日益坚实的桥梁，逐渐形成了独特的跨学科研究范式，广泛应用在图情学、管理学、文学、历史学、艺术学等多个学科。近年来，国内外大量数字人文研究机构成立，领域学者群体规模迅速扩大，相关研究成果快速增长，使得数字人文研究逐步从概念界定和框架设计落脚到实证研究[1]。浩瀚的传统典籍成为数字人文实践研究的重要对象，被誉为“一方之全史”的地方志以十分之一的占比位居古籍大宗，体系完备、类目繁多，有舆地、河渠、建置、文教、赋役、职官、人物、物产、艺文等内容，全面记载了特定时空下自然、社会、经济、政治、文化等各个方面的情况[2]，承担着文化传承、智慧延续、经验启示的历史使命，是中国乃至世界重要的文化遗产宝库[3]。如何借助数字人文的技术优势，深度挖掘和利用地方志文献内容，充分发挥其存史、资政、教化、兴利等重要作用，是一项值得长期关注的课题。本文以《方志物产》山西分卷为研究语料，面向其中蕴含的物产用途实体，基于Bi-LSTM、Bi-LSTM-CRF、BERT、Siku-BERT等4种深度学习模型实现自动识别模型构建和实体识别效果比较，为以地方志为代表的大规模典籍的文本挖掘与开发利用提供借鉴。

1 相关研究综述

命名实体识别（Named Entities Recognition）作为数字人文研究的重要环节，承担着从文本中自动抽取具有特定意义实体（包括人名、时间、地名、机构名等专有名词）的关键任务，发挥着基础性作用。在面向中文古籍的数字人文研究中，基于特定语料的命名实体识别研究持续时间长、研究成果多。

早期的命名实体在基于规则的基础上开展起来：朱晓[4]以编年体《明史本纪》为例，对人名实体进行了自动识别，取得了较好的效果；衡中青[5]以《方志物产》

广东分卷为研究对象，进行了引书名和别名的自动抽取，正确率分别为72.88%、71.60%；朱锁玲[6]面向《方志物产》广东、福建、台湾三省语料，开展了地名的识别，正确率达到了63.38%；刘士纲[7]面向《清实录》语料，采用统计与规则相结合的方式进行了人名识别。随着研究深入和技术提升，基于条件随机场（Conditional Random Field，CRF）的方法大大提升了识别效果：汪青青[8]对先秦文献《春秋左传》中的人名识别开展了实验，开放测试准确率达到了92.48%；肖磊[9]对《左传》中的地名进行了自动识别实验，正确率达到了94.59%；李章超等[10]有效抽取了《左传》文本中的战争事件；黄水清等[11]基于先秦语料库，分别使用条件随机场和最大熵模型对地名进行了识别，验证CRF模型识别效果较好；王铮[12]以《三国演义》为研究语料，自动抽取了地名实体，准确率为99.16%；叶辉等[13]基于融合多特征的CRF模型实现了中医古籍《金匮要略》中的症状药物实体抽取；李娜[14]以《方志物产》山西分卷为例，对其中蕴含的人名、地名、别名、引书名和用途名等实体进行了自动识别。近年来，深度学习模型发展迅速，呈现出较好的应用态势：李成名[15]基于LSTM-CRF模型对《左传》中蕴含的人名和地名开展了实验，识别超过到82%；徐晨飞等[16]运用Bi-RNN、Bi-LSTM、Bi-LSTM-CRF、BERT等模型，自动抽取了《方志物产》云南分卷中人名、别名、地名和引书名等实体，取得了较好的实验效果；李焕[17]面向中医古籍本文，使用BERT-BiLSTM-CRF模型对中医术语进行了识别，有效提升了识别效果；刘忠宝等[18]将BERT和LSTM-CRF模型应用到《史记》中历史事件的自动抽取中，*F*值为82.3%；杜悦等[19]通过7个深度学习模型在25本典籍语料中历史事件的抽取实验，证明了深度学习模型在大规模古籍文本整理的适用性；崔竞烽等[20]通过实验论证了BERT模型在菊花古典诗词中的时间、地点、季节、花名、花色、人物、节日等实体的识别效果较好；黄水清等[21]对比了CRF、Bi-LSTM、Bi-LSTM-CRF模型在《论语注疏》《毛诗正义》《春秋左传正义》三部典籍中引书名的识别效果，验证了深度学习模型的整体表现明显优于CRF模型；范涛等[22]基于《人民日报》语料库和中文推特多模态数据集预训练了BiLSTM-attention-CRF模型，并迁移至地方志文本中开展多模态识别实验，具有一定的优势；任常青[23]运用CRF、Bi-RNN、Bi-LSTM-CRF模型对雄安县地方志中记载的七大类实体进行了自动抽取，发现融合机器学习与深度学习模型在大规模古籍文本的深度挖掘中具有更好的表现；刘江峰等[24]以“前四史”和《左传》为语料，对比了Bert-base、guwenBert、sikuBERT、SikuRoBERT等预训练深度学习模型的实体识别效果，验证了sikuBERT模型的优越性。

笔者通过梳理发现，命名实体识别研究经历了基于规则、统计、深度学习三个阶段的发展，识别模型越来越智能化，识别效果逐步提高，识别对象以人名、地名、事件、事件名为主，也涉及中药名、方剂名、物产别名、引书名等实体类型，但对于用途实体类型的识别研究成果较少，程度较浅，有一定的探索和深化空间。

2 《方志物产》语料简介

本文的研究对象《方志物产》，是一套汇集方志中物产类目的专题资料，藏于南京农业大学图书馆。20世纪50年代，在著名农史学家万国鼎先生的主持下，数十名专业人员奔赴全国40多个大中型城市，从100多个文史单位保存的7 200余部地方志中，手工摘抄了物产部分资料，根据地区和时间顺序编纂成册，覆以红皮，俗称农史学界的“红本子”[25]。

首先，横跨地域范围广，囊括青海省、新疆维吾尔自治区、西藏自治区、台湾省在内的所有行政区域；其次，纵向时间区间长，从宋熙宁九年（1076年）的《长安志》至民国三十八年（1949年）的《定西县志》，持续时间近900年；再次，所载物产种类多，全文共449卷、3 600余万字，记载了植物、动物、货物（天然产矿物和人工造货物）、微生物等153万余条物产信息；最后，来源志书类型全，包括全国总志、省志、府志、州志、县志、区志、村志、祠庙志、乡土志、山水志、边关志等多种类型。自编纂以来，《方志物产》因其独特的价值受到了学界的高度重视，国内外众多学者前往查询，为区域发展、学术研究等提供了丰富的资料支撑。

在《方志物产》的记载体例中，主要内容是物产名称和对应的描述信息，其中，描述信息中主要记述了物产的别名、生长环境、生物学特征（大小、颜色、形状等）、引用的其他典籍名称、相关的历史人物、产地、产量、价格、用途等内容，有全有缺、有详有略，具体语料样例如表1所示。

关于物产的描述信息中蕴含的人名、地名、别名、引书名等实体识别研究，衡中青等[26]、朱锁玲等[27]基于规则的实验开展了引书名、地名的识别，李娜[28]基于CRF

表1 《方志物产》语料样例

序 号	物产名称	描述信息	描述内容
1	何首烏	產九華山者良於他處	产地
2	芝麻	有黑白二種黑者即胡麻也久服可以延年	颜色、别名、用途、品种
3	石蘭	一莖一葉一花色紫生峭壁上	外形、颜色、生长环境
4	瞿麥	爾雅作蘧麥	引用名、别名
5	薺	爾雅曰大薺薪冥又名靡草俗稱薺菜又歲欲豐甘草先生即薺也詩曰其甘如薺蘇軾謂取薺莖肥大者洗淨米屑拌之入油少許覆碗蒸爛味等淳熬信然花置席下可辟蟲子主明目	引用名、别名、人物、用途、用法
6	小麥	年產十萬石	产量
7	山羊皮	一萬二千張所出皮毛，除供本縣用外，餘皆運銷於包頭	产量、贸易
8	織其	年產織其一萬二千餘斤，每百斤值洋二元	产量、价格
9	石灰	常産	无
10	氈毯	以牛馬羊毛作之	制作原材料

模型的实验自动抽取了人名、地名、别名、用途名、引书名，徐晨飞[29]基于深度学习模型的实验进行了人名、地名、别名、引书名的识别，效果对比显著。针对其中物产的用途实体识别仅有李娜面向《方志物产》山西分卷基于CRF模型的研究，识别效果在70%左右[28]。

本文所谓物产用途是指物产的功能，如黄豆可以榨油、造酱、制成豆腐，黄精具有补中、益气、轻身延年的功用；桦木可以制成刀靶和酒器、装饰弓箭；桑，不仅可以饲蚕，也可以制弓、编筐，还可以治疗咳嗽。通过深度梳理物产的用途实体，可以明确一个物产具有哪些用途，哪些物产具有相同或者相似的用途，以便更加全面而深刻地认识物产，有助于探索更加科学的途径实现物产价值的开发利用。经过梳理，《方志物产》所载的物产用途主要分为以下方面：饮食方面，作为食物充饥；药用方面，作为药材进行疾病防治；经济方面，服务生产生活之用；民俗方面，祈福、辟邪等功能；制毒方面，有毒成分，使用可以致病、致死等。物产的用途语料样例如表2所示。

通过阅读和分析《方志物产》中用途实体的分布特征和记载规律，例如：物产的用途实体分布范围较广且不集中，有的紧跟物产名，如艾的描述中“艾葉療一切鬼氣”；有的出现在物产结构部位的后面，如扁豆的描述中“莢可蔬粒可羹”；有的伴随物产形状表述，如蓖麻的描述中“色青而黑用以榨油”；也有单独记载的，如柏的描述信息“可作棺祭器”；等等。物产的用途实体记载具有一定的规律性，如常以“作、可、治、疗、以”等作为边界词进行定位，但仍有大量规则之外的多样性表达形式。这些文本特征都为用途实体抽取带来了

表2 含用途实体的物产语料样例

序 号	物 产	描述信息	用 途
1	艾	園野皆產然不如蘄州之佳郭璞雲艾葉療一切鬼氣	民俗：疗一切鬼气
2	白菜	一名菘有黃芽一種更佳通和腸胃消酒毒	药用：通和肠胃、消酒毒
3	柏	可作棺祭器	经济：作棺、祭器
4	鮀鮀	食之殺人	制毒：致死
5	蓖麻	俗名大麻色青而黑用以榨油	经济：榨油
6	扁豆	花如小蛾翅尾俱全莢可蔬粒可羹	饮食：蔬菜、做羹
7	波羅花	最毒食之令人顛狂不可植	制毒：致病
8	小豆	有赤白二種赤者可避瘟	药用：避瘟防疫

困难，导致前期仅依赖规则和统计的识别结果中混杂着非用途实体、部分用途实体的遗漏或者实体字符长度有误等现象。在人工标注语料的基础上，深度学习模型可以灵活地结合上下文的语义关系，充分学习文本的结构特征，通过自动识别模型构建和识别效果评价机制，优化目标实体抽取的效率和精度，提升方志类古籍文本挖掘研究进程。

3 模型介绍

3.1 Bi-LSTM、Bi-LSTM-CRF循环类深度学习模型

循环类深度学习模型主要是运用具有信息保存

能力的循环神经网络（Recurrent Neural Network，RNN），在自然语言处理的过程中通过对长特征向量预测当前输出，以解决序列标注问题[30]。但由于RNN依靠单一隐藏层的记忆结构过于简单，随着输入序列长度的增加会出现梯度消失的问题，从而限制了模型处理长输入序列的效果提升。长短时记忆网络（Long Short-Term Memory，LSTM）引入细胞状态记忆单元、输入门（input gate）、忘记门（forget gate）、输出门（output gate）实现信息的存储[31]，然而无法实现从后往前的信息编译问题。

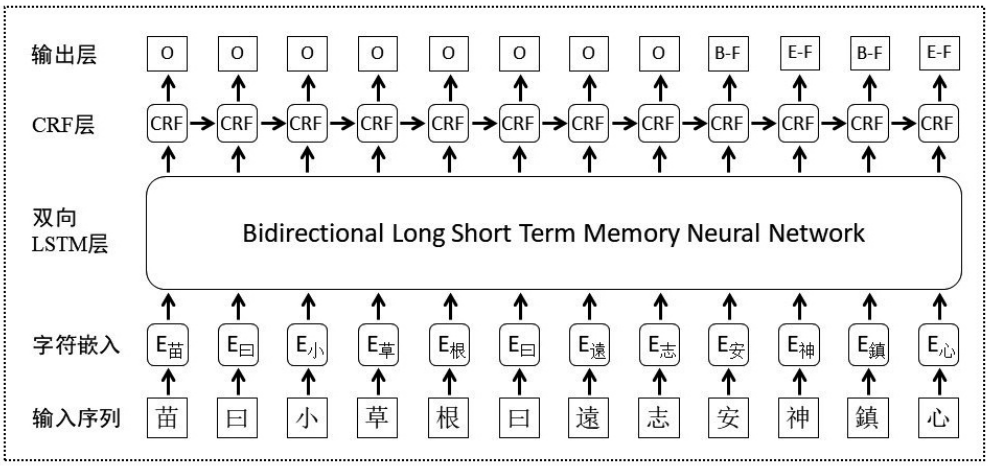
双向长短时记忆网络（Bi-Directional Long Short-Term Memory，Bi-LSTM）将向前的LSTM和向后的LSTM相结合，在保留输入门、输出门和忘记门的基础上，解决了文本上下文信息表示的问题[32]。面向给定的语句，先将其中蕴含的每个词t处理成一个长度为d的向量，使用模型计算出t的左边上下文向量和右边上下文向量，表示为向量$h_t=[\overrightarrow{h_t};\overleftarrow{h_t}]$，层级关系输入的文本序列，经过双向长短时记忆层预测，层级关系得到输出的标注序列。本文运用“S、B、I、E、O”标签机制，其中S表示用途实体本身，B表示用途实体的起始字，I表示用途实体的中间字，E表示用途实体的结束字，O表示用途实体以外的字。以《方志物产》山西分卷中“紫背者良发汗利湿”这句物产描述语料为输入序列样例，对应的输出标注标签为“O”“O”“O”“O”“B-nf”“E-nf”“B-nf”“E-nf”，其中“nf”为用途实体类型标注。可见，双向长短时记忆网络在解决长距离依赖问题的基础上，同时兼顾了整个语料上下文的序列信息，提高了深度学习模型的识别性能。

为了进一步提升循环类深度学习模型的识别性能，本研究将双向长短时记忆网络与线性条件随机场模型相结合，形成Bi-LSTM-CRF模型，是目前解决序列标注问题的主流方法，既保存了文本序列中上下文信息，又考虑到句子中标签之间的转移关系，有效解决了序列标注中标记偏置问题[33]。首先将输入文本序列进行词向量化处理，其次利用双向LSTM层获得的上下文特征和CRF层输入的语句级别标记序列，最后经过CRF层对全局进行状态转移概率计算，提取实体之间的依赖关系，从而实现预测标签信息，并结合动态规划的Viterbi算法找到整个句子的最佳标签序列。

在面向《方志物产》语料的循环类实体识别实验中，以基于Bi-LSTM-CRF的模型为例，使用标签机制对物产“遠志”的描述信息“苗曰小草根曰遠志安神鎮心令人多記”进行标注（见图1），其中“安神”“鎮心”“令人多記”均为物产的用途实体，可以看出，实体标签之间存在紧密的逻辑关系，并且同时受到上下文语义以及标签的双重影响。

3.2 BERT、Siku-BERT注意力机制类深度学习模型

注意力机制（Attention Mechanism）是在机器学习模型中嵌入的一种特殊结构，实现自动学习和注意力分布加权计算输入数据对输出数据的贡献大小，最早应用于图像分类领域，随后引入自然语言处理领域。

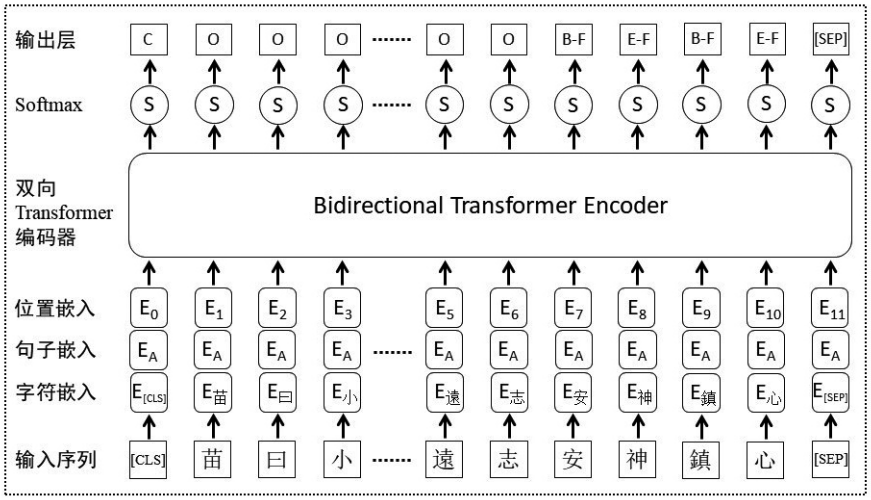


苗曰小草根曰遠志【F安神】【F鎮心】【F令人多記】

图1 基于Bi-LSTM-CRF的《方志物产》用途实体识别模型原理图

BERT（Bidirectional Encoder Representations from Transformers）模型以多层Transformer结构为主要框架，其强大的特征提取能力，有效解决了语料长依赖问题，相对于循环类神经网络模型具有明显优势[34]。BERT模型在预训练（pre-training）阶段，将句子中各个词或字的原始向量作为输入部分，通过对应的位置嵌入（position embedding）、分割嵌入（segment embedding）和Token嵌入（token embedding）的求和构造输入表示（input representation），在当前句子与上一个句子和下一个句子的Token位置，分别嵌入[CLS]和[SEP]标记进行句子分割，而输出部分则是融合了全文语义特征后的文本中词或字的标记向量，利用遮盖和预测方法学习两个句子之间的关系。

在目前面向中文古籍的BERT预训练模型中，guwenBERT是基于殆知阁简体中文古籍文献训练的，Bert-Base-Chinese和RoBERTa模型则是基于简体与繁体相融合的中文维基百科数据进行训练的，Siku-BERT是基于《四库全书》繁体版的预训练模型。从语料相似和功能需求的角度，本文选择基于繁体古籍的预训练模型Siku-BERT开展具有监督的预训练-参数微调范式实验，以验证注意力机制模型在迁移过程中的语料适用性。Siku-BERT迁移到《方志物产》山西分卷的实体抽取任务时，在双向Transformer模型的编码器基础上，通过随机遮蔽字符，使得模型以自监督的方式从前后两个方向同时预测被遮蔽字符，从而更加有效地学习到语料的文法、句法、语言风格等特征[35]，如图2所示。



苗曰小草根曰遠志【F安神】【F鎮心】【F令人多記】

图2 基于Siku-BERT的《方志物产》用途实体识别模型原理图

4 实验开展

4.1 语料选择与预处理

《方志物产》山西分卷共13册、全文约43万字，始于明成化二十一年（1485年）的《山西通志》，截止于民国二十九年（1940年）的《榆次县志》，记载了455年间山西境内产出的植物、动物、货物（包括天然的矿产和人造的商品）等51 545条物产的分类、名称以及描述信息。经统计，该卷共包含316本志书，其中，根据时间维度划分，明代46本，清代237本，民国时期33本；根据地区划分，平阳府以52本居首，其次是太原府36本，潞安府31本位居第三；根据类型划分，有通志6本，府州志41本，县志258本，乡土、山川、偏关等志11本。

如前文所示，在物产的描述信息中，蕴含着大量的关键要素，如人名、地名、引书名、别名、用途名等实体，由于对人名、地名、引书名和别名识别的已有研究较多且取得了较好的识别效果，本文主要面向关注度较低的物产用途实体进行识别，以拓展信息抽取的实体类型，验证深度学习模型在更多类型实体抽取中的效果。物产用途实体主要反映了物产在生活、医疗、民俗等方面的作用，如建造房屋、治疗疾病、祈福辟邪等。研究物产的用途，不仅可以展现古人在物产利用方面的智慧与思考，也为当今物产的开发利用提供经验借

鉴和方法路径。

由于《方志物产》中物产的描述信息不全，导致一部分物产没有描述信息。因此，在语料的预处理过程中，首先筛选出含有描述信息的物产数据，再对物产的描述信息进行人工标注，以“F”代表用途实体的标识符，用“【】”表示用途实体的左右边界，如物产“柏子仁”的描述信息为“氣味清香補脾養心潤腎滋肝”，标注后的语料为“氣味清香【F補脾】【F養心】【F潤腎】【F滋肝】”。

4.2 实验方法与设置

本文使用“B、I、E、O”四词位的标注集，如物产“遠志”的描述信息为“苗曰小草根曰遠志安神鎮心令人多記”，经过标注处理后的结果见表3。另外，为了适应《方志物产》的语言特性，深度学习模型的输入特征定义为字符向量，通过Word2Vec[36]获得，实现文本潜在语义的自动搜索，降低特征模板对词语长度、出现频次、左右边界词等语料内外部特征的依赖。还有，本文采用十次交叉法验证不同训练语料下的模型识别效果，已取得更加科学精确的实验效果。即以整句为单位，将人工标注后的语料进行随机乱序排列，每次选取其中9份作为训练语料，剩余一份作为测试语料，分别对4种深度学习模型进行多次性能测试，寻找最优模型。

表3 《方志物产》物产用途标注语料样例

序 号	词 语	标 记	序 号	词 语	标 记
1	苗	O	9	安	B
2	曰	O	10	神	E
3	小	O	11	鎮	B
4	草	O	12	心	E
5	根	O	13	令	B
6	曰	O	14	人	I
7	遠	O	15	多	I
8	志	O	16	記	E

在运算过程中，神经网络模型需要进行大量的并行计算，一般的中央处理器无法满足，本文所进行的神经网络训练实验使用了高性能的NVIDIA Tesla P40图形处理器，充分保障吞吐量和相应速度的需求，具体参数如表4所示。

表4 实验超参数设置表

超参数	值
Bi-LSTM/Bi-RNN层数	2
Hidden size	256
Learning rate	0.001
Batch-size	64
Dropout	0.5
Clip gradient	5

BERT模型因其语言模型和特征抽取架构的独特性，运算时需要更大的空间支持，与传统深度学习模型的参数设置有一定差异性，具体如表5所示。

表5 BERT模型的实验超参数设置表

超参数	值
BERT	2
Hidden size	128
Learning rate	2e-5
Batch-size	32
Train-epochs	3

4.3 实验结果

面向《方志物产》山西分卷语料，运用Bi-LSTM、Bi-LSTM-CRF、BERT、Siku-BERT等4种深度学习模型对文本中蕴含的物产用途实体进行自动抽取，对比不同语料、不同模型的实验效果。本研究采用准确率P、召回率R和调和平均数F作为评级指标，具体的计算公式如下。

$$P=\frac{\text{识别正确的实体}}{\text{识别正确的实体+识别错误的实体}}\times 100\% \quad (1)$$

$$R=\frac{\text{识别正确的实体}}{\text{识别正确的实体+未被识别的实体}}\times 100\% \quad (2)$$

$$F=\frac{(\beta^2+1)\times P\times R}{(\beta^2\times P)+R}\times 100\%=\frac{2\times P\times R}{P+R}\ (\beta=1) \quad (3)$$

经过语料标注和相应的深度学习模型构建，运用十次交叉法对模型性能进行测试的结果如表6所示。

可以看出，在未使用任何人工提供特征进行构建特征模板的情况下，4种深度学习模型的测试结果呈现出一定的差异性。从召回率的结果看，BERT模型达到了88.62%，Siku-BERT模型则达到了89.74%，验证了深

表6　4种深度学习模型的十次交叉测试结果　%

Bi-LSTM				Bi-LSTM-CRF			BERT			Siku-BERT		
序　号	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
1	63.64	64.62	64.12	60.00	60.00	60.00	82.09	84.62	83.33	77.78	86.15	81.75
2	72.44	74.80	73.60	78.05	78.05	78.05	81.30	81.30	81.30	76.92	81.30	79.05
3	73.81	72.09	72.94	74.38	69.77	72.00	86.15	86.82	86.49	83.58	86.82	85.17
4	55.26	69.42	61.54	63.91	70.25	66.93	77.69	83.47	80.48	79.40	89.26	84.05
5	65.15	66.67	65.90	64.19	73.64	68.59	75.89	82.95	79.26	75.89	82.95	79.26
6	69.12	73.44	71.21	66.91	72.66	69.66	75.54	82.03	78.65	74.83	83.59	78.97
7	65.00	73.98	69.20	69.57	65.04	67.23	82.58	88.62	85.49	79.70	86.18	82.81
8	64.62	71.79	68.02	68.75	75.21	71.84	81.45	86.32	83.82	80.15	89.74	84.68
9	70.68	73.44	72.03	72.87	73.44	73.15	79.23	80.47	79.84	76.81	82.81	79.70
10	73.20	74.67	73.93	67.72	71.33	69.48	81.65	86.00	83.77	82.10	88.67	85.26
平均值	67.29	71.49	69.25	68.63	70.94	69.69	80.36	84.26	82.24	78.72	85.75	82.07

度学习模型在《方志物产》语料用途实体识别中的适用性。相较于前期基于CRF模型的物产用途实体识别结果，深度学习模型全面提升了识别效果，更加凸显了其优越性。另外，还可以发现以下现象。

（1）相较于Bi-LSTM模型，Bi-LSTM-CRF模型识别结果有了明显提升。说明引入CRF层后，增强了序列标注问题的处理能力，将上下文特征与规则和统计方法充分结合，有助于提升古方志实体识别的效果。

（2）BERT模型较Bi-LSTM和Bi-LSTM-CEF模型总体上有显著提升，证明基于注意力机制的多层双向Transformer架构的预训练模型在大规模古方志语料实体抽取中的突出性。

（3）Siku-BERT模型和BERT模型的识别性能相似，均取得了较为显著的结果，BERT模型的*P*和*F*结果略优于Siku-BERT，而Siku-BERT的*R*值略高于BERT，验证了基于《四库全书》开发的Siku-BERT模型迁移至方志古籍语料实体识别可行性。

4　原因分析

经过4种深度学习模型的识别结果与人工标注语料的详细对比，发现在语料标注、语料规模、语料特征等方面存在不足之处，后续经过语料规模的扩展以及人工标注的完善，实验效果还有提升的空间。

（1）人工标注存在漏标现象。主要表现为部分“可食”用途的漏标，在阅读和标注的过程中，在“可食”用途与其他更凸显的用途共同出现在一条语料中时，容易忽略对“可食”用途的标注，如“菜籽”的描述信息为“隨砑穀而種初夏收割花黃葉亦可食其籽用以榨油人多食之”，在标注时仅标注了菜籽“用以榨油”，而漏标了叶亦“可食”。再如“長松”的描述信息“能治大風氣味芳烈採之可作湯常服亦名仙茅唐時即著名宋僧延一舊志云出東西兩臺”，则仅标注出了“能治大風”，漏标了“可作湯”这个实体。

（2）结构特征呈现无边界且连续表达的情况。在描述一个物产用途时，经常会将其具有数个用途连续记载，每个用途实体之间没有边界词加以区分，这对于人工处理十分棘手，计算机自动分词则更加困难。例如，物产“青蒿”的描述信息为“處處生之春夏採莖葉同童便煎退骨蒸劳熱生搗絞汁却心疼熱秋黃冬採根實實須炒治風癆疥瘙虛煩盜汗開胃明目辟邪殺虫”，其中就有“風癆、疥瘙、虛煩、盜汗、開胃、明目、辟邪、殺虫”等多种用途名称的无边界连用。

（3）实验语料规模较小。基于深度学习的模型减少了人工提取特征的依赖，更适合在大数据的环境下，开展对大规模语料特征的自动学习。但本实验仅使用了《方志物产》山西省的语料，尽管实验结果证明了深度学习模型的优越性，但语料规模远达不到其对大规模语料的需求。随着多省语料的逐步整合，语料规模日益扩大，深度学习模型的识别效果也将随之不断提高。

5　应用场景

经过分析物产用途实体的识别结果可以发现，物

产的主要用途集中在食用和药用两大方面。其中，物产的食用分为非加工品和加工品两种类型，非加工品主要是指自然界产出的瓜、果类物产（如苹果、西瓜等），而加工品则是通过腌、煮、蒸、炒、炸等方式进行加工处理的物产（如咸菜、麻油等）。物产的药用体现在预防、治疗、致毒三个方面，关系人类健康的方方面面，记载较为翔实。本章选取了物产用途中的药用方面，通过物产与药用的关联关系，展示识别结果的相关应用。

5.1 物产的药用关系网络构建

基于社会网络分析技术，构建物产与药用的关系网络，得到由195种物产名、323种药用名、429条连线所组成的网络。为了更加清晰地展示，本文抽取了其中的最大联通子网络的局部图（见图3）。其中，圆圈代表物产名，方框代表药用名，颜色深浅代表相连顶点的不同数量，连线代表相连的物产具有某种药用价值。通过

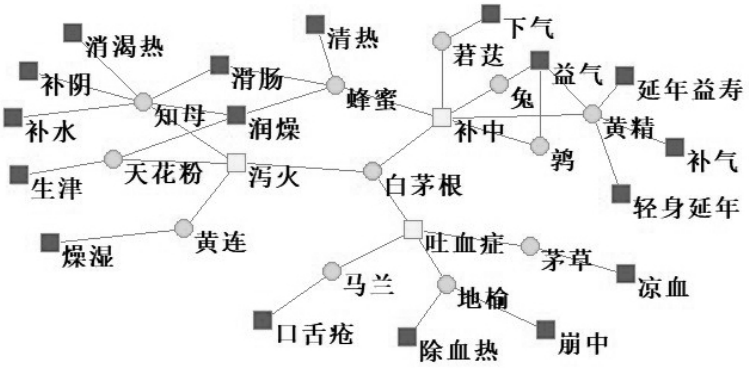


图3 物产的药用关系网络图

该图可以发现某个物产具有什么药用价值，如物产“马兰”具有治疗“吐血症”和“口舌疮”的功能；还能发现具有相同药用价值的物产有哪些，如具有“泻火”作用的物产有“知母”“天花粉”“黄连”和“白茅根”。

5.2 探寻物产的药用部位

根据用途实体所处的上下文语义，可以快速获取具有特定药用价值的具体部位，有助于药材的深度开发利用。经分析，物产的药用部位分布较广，有单一部位可用物产，也有多部位可用物产，甚至还有全身可用物产；不同物产的药用部位差异性较大，同一个物产不同部位的用途也可能不同。总体而言，植物的药用部分主要有叶、花、根、茎、实、皮、汁等，动物的药用部分主要有肉、卵、血、屎、涎等，具体案例如表7所示。

6 结语

本文使用Bi-LSTM、Bi-LSTM-CRF、BERT、Siku-BERT等4种深度学习模型，针对《方志物产》山西分卷文本中蕴含的物产用途实体进行了自动识别。实验结果显示，4种深度学习模型的*P*、*R*、*F*值均高于笔者此前基于CRF构建的抽取模型[28]，特别是BERT模型和Siku-BERT模型的识别性能较为突出，召回率*R*的最高值分别为88.62%和89.74%。本研究拓展了物产用途实体抽取的应用类型，有效提升了实体识别的效果，同时验证了深度学习模型在方志类古籍整理中的可行性以及预训练模型的可迁移性。

当然，若要进一步提升深度学习模型的识别性能，构建更加丰富、立体、全面的物产用途网络，后续研究不仅要不断扩大语料规模和完善语料标注，从单一省份扩展到多个省份甚至一个地区乃至全国范围，覆盖更多、更全的语料特征，不断优化深度学习模型，为方志类古籍的开发利用提供数字人文领域的解决方案，为更多特色馆藏文献的整理挖掘提供借鉴。

参考文献

[1] 赵薇. 数字时代人文学研究的变革与超越——数字人文在中国 [J]. 探索与争鸣，2021（6）：191-206，232-233.

[2] 来新夏. 方志学概论 [M]. 福州：福建人民出版社，1983：6-32.

[3] 林衍经. 方志学综论 [M]. 上海：华东师范大学出版社，2008：7-12.

[4] 朱晓. 古汉语编年体的人名实体识别与词性标注 [D]. 上海：复

表7 物产的药用部位示例

物产名	描述信息	药用部位	药用名
艾	園野皆產然不如蘄州之佳郭璞雲艾葉療一切鬼氣	叶	疗一切鬼气
大剪	花如梅大剪葉皺小剪相反皆用根能破血下氣	根	破血、下气
鴿	性溷雌常乘雄能傳書故又名飛奴好寄人字下肉美能解諸藥毒卵能佐筵小兒食之永不出痘	肉	解诸药毒
		卵	小儿食之永不出痘
狗尾草	苗葉似粟殼而穗黃形似狗尾即莠草也莖能治目痛故又名光明草	茎	治目痛
家鵝	其血治噎嗝反胃症	血	治噎、嗝、反胃
蚯蚓	其屎能治蠍螫	屎	治蝎螫
蝸牛	其涎可治蠍螫	涎	治蝎螫
馬兜鈴	蔓生實如鈴體輕去筋膜取子用能降肺氣治熱欬嗽	子	降肺气、治热欬嗽
牡丹	生山谷中皮主通經除冷	皮	通经、除冷
金絲	搗汁可治耳病有翠雲喜生陰地蒼翠如雲	汁	治耳病
紅果	一名林禽似沙果而小味甘溫下氣消渴	果	下气、消渴
紅花	古名紅藍花活血潤燥治經閉症	花	活血、润燥、治经闭症

旦大学，2012.

[5] 衡中青. 地方志知识组织及内容挖掘研究 [D]. 南京：南京农业大学，2007.

[6] 朱锁玲. 命名实体识别在方志内容挖掘中的应用研究 [D]. 南京：南京农业大学，2011.

[7] 刘士纲. 《清实录》人名撷取自动化 [D]. 台北：台湾大学，2012.

[8] 汪青青. 先秦人名识别初探 [J]. 文教资料，2009（18）：202-204.

[9] 肖磊. 先秦地名知识库构建 [D]. 南京：南京师范大学，2010.

[10] 李章超，李忠凯，何琳. 《左传》战争事件抽取技术研究 [J]. 图书情报工作，2020，64（7）：20-29.

[11] 黄水清，王东波，何琳. 基于先秦语料库的古汉语地名自动识别模型构建研究 [J]. 图书情报工作，2015（12）：135-140.

[12] 王铮. 基于CRF的古籍地名自动识别研究 [D]. 南宁：广西民族大学，2008.

[13] 叶辉，姬东鸿. 基于多特征条件随机场的《金匮要略》症状药物信息抽取研究 [J]. 中国中医药图书情报杂志，2016，40（5）：14-17.

[14] 李娜. 社会网络分析视角下方志古籍知识组织研究 [D]. 南京：南京农业大学，2017.

[15] 李成名. 基于深度学习的古籍词法分析研究 [D]. 南京：南京师范大学，2018.

[16] 徐晨飞，叶海影，包平. 基于深度学习的方志物产资料实体自动识别模型构建研究 [J]. 数据分析与知识发现，2020，4（8）：86-97.

[17] 李焕. 基于深度学习与主动学习的中医术语识别研究 [D]. 北京：北京工业大学，2019.

[18] 刘忠宝，党建飞，张志剑. 《史记》历史事件自动抽取与事理图谱构建研究 [J]. 图书情报工作，2020，64（11）：116-124.

[19] 杜悦，王东波，江川，等. 数字人文下的典籍深度学习实体自动识别模型构建及应用研究 [J]. 图书情报工作，2021，65（3）：100-108.

[20] 崔竞烽，郑德俊，王东波，等. 基于深度学习模型的菊花古典诗词命名实体识别 [J]. 情报理论与实践，2020，43（11）：150-155.

[21] 黄水清，周好，彭秋茹，等. 引书的自动识别及文献计量学分析 [J]. 情报学报，2021，40（12）：1325-1337.

[22] 范涛，王昊，陈玥彤. 基于深度迁移学习的地方志多模态命名实体识别研究 [J]. 情报学报，2022，41（4）：412-423.

[23] 任常青. 数字人文视角下县志作物类物产实体识别研究——以雄安县志为例 [J]. 信息与电脑（理论版），2022，34（1）：74-76.

[24] 刘江峰，冯钰童，王东波，等. 数字人文视域下SikuBERT增强的史籍实体识别研究 [J]. 图书馆论坛，2022，42（10）：61-72.

[25] 包平，李昕升，卢勇. 方志物产史料的价值、利用与展望——以《方志物产》为中心 [J]. 中国农史，2018，37（3）：117-126.

[26] 衡中青，刘竟，侯汉清. 《方志物产》引书挖掘及分析研究——以《岭南丛述》（物产）为例 [J]. 中国农史，2007（3）：132-139.

[27] 朱锁玲，包平. 方志类古籍地名识别及分析研究——以《方志物产》（广东分卷）为例 [J]. 图书馆论坛，2012，32（4）：171-176.

[28] 李娜. 面向方志类古籍的多类型命名实体联合自动识别模型构

建[J]. 图书馆论坛，2021，41（12）：113-123.

[29] 徐晨飞. 数字人文视域下方志物产知识库构建研究——以《方志物产》云南卷为例[D]. 南京：南京农业大学，2020.

[30] 邱锡鹏. 神经网络与深度学习[EB/OL]. [2022-07-05]. https://nndl.github.io/nndl-book.pdf.

[31] GREFF K，SRIVASTAVA R K，KOUTNÍK J，et al. LSTM：a search space odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems，2016，28（10）：2222-2232.

[32] 杨培，杨志豪，罗凌，等. 基于注意机制的化学药物命名实体识别[J]. 计算机研究与发展，2018，55（7）：1548-1556.

[33] HUANG Z H，XU W，YU K. Bidirectional LSTM-CRF models for sequence tagging[J/OL]. arXiv Preprint，arXiv：1508.01991[2022-12-20]. DOI：https://doi.org/10.48550/arXiv.1508.01991.

[34] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding[J/OL]. arXiv Preprint，arXiv：1810.04805[2022-12-20]. DOI：https://doi.org/10.48550/arXiv.1810.04805.

[35] 胡昊天，张逸勤，邓三鸿，等. 面向数字人文的《四库全书》子部自动分类研究——以Siku BERT和Siku Ro BERTa预训练模型为例[J/OL]. 图书馆论坛：1-16[2022-07-05]. http://kns.cnki.net/kcms/detail/44.1306.G2.20211017.1823.002.html.

[36] RONG X. Word2vec parameter learning explained[J/OL]. arXiv Preprint，arXiv：1411.2738[2022-12-20]. DOI：https://doi.org/10.48550/arXiv.1411.2738.

作者简介

李娜，女，1985年生，博士，讲师，研究方向：数字人文、资源管理与利用，E-mail：rwlina@njfu.edu.cn。

Construction of Automatic Recognition Model of Function Entities in *Local Chronicles*: *Produce* Based on Deep Learning

LI Na

(Faculty of Social Sciences and Humanities Nanjing Forestry University, Nanjing 210037, P. R. China)

Abstract: Taking the local chronicles as the research corpus, based on the manually labeled corpus, we use four deep learning models such as Bi-LSTM, Bi-LSTM-CRF, BERT and Siku-BERT to carry out experiments, and then use the accuracy rate P, recall rate R and F-value as test indicators to compare and analyze the recognition performance of the models, so as to promote the mining and utilization of product knowledge. The experimental results show that: Compared with the previous model based on CRF, the overall performances of the four deep learning models have been significantly improved; The best R-values of Bi-LSTM, Bi-LSTM-CRF, BERT and Siku-BERT are 74.80%, 78.05%, 88.62% and 89.74% respectively; The recognition effects of attention mechanism deep learning models such as BERT and Siku-BERT are better than those of cyclic deep learning models such as Bi-LSTM and Bi-LSTM CRF. Although there is still much room for further optimization of the model performance due to the structural characteristics of the local chronicle ancient books, the imperfect manual annotation of the corpus, the small scale of the corpus and other factors, the deep learning models show certain superiority in the content mining of local chronicles, and the migration and application of the pre-training model between different corpora are feasible.

Keywords: Deep Learning; *Local Chronicles*: *Produce*; Named Entity Recognition; Digital Humanities; Function Entities

（收稿日期：2022-11-08）