古农书翻译与知识组织研究

吴梦成 王东波 黄水清

(南京农业大学 领域知识关联研究中心/信息管理学院,江苏 南京 210095)

【摘 要】在探寻古农书知识的深层脉络中,文章从跨学科视角开发了一种创新的古农书翻译与知识图谱构建方法。该方法不仅有利于深化对古农书的理解,也为古农书的文化传承与现代应用提供了新的思路。首先通过训练和微调古代汉语到现代汉语的机器翻译模型,完成古农书的翻译并构建古农书平行语料库以辅助人工标注与古农书知识图谱构建。接着,本文通过知识标注、实体抽取、关系抽取等数据挖掘技术,对古农书的非结构化知识进行系统梳理与组织。最终,构建的知识图谱不仅仅是对古农书知识的再现,更是对这些知识进行深入探索与应用的工具。研究结果揭示了所提方法在促进古农书知识理解、应用与传播方面的有效性以及跨学科方法在古农书研究中的重要性。

【关键词】古农书;机器翻译;知识标注;知识图谱

【中图分类号】S-09;K207 【文献标志码】A 【文章编号】1000-4459(2024)02-0052-13

Research on the Translation and Knowledge Organization of Ancient Agricultural Books

WU Mengcheng WANG Dongbo HUANG Shuiqing

(*Research Center for Correlation of Domain Knowledge/School of Information Management, Nanjing Agricultural University, Nanjing 210095*)

Abstract: In exploring the deep context of ancient agricultural book knowledge, this paper develops an innovative method of ancient agricultural book translation and knowledge graph construction from an interdisciplinary perspective. This method not only helps to deepen the understanding of ancient agricultural books, but also provides new ideas for the cultural inheritance and modern application of ancient agricultural books. First of all, by training and fine-tuning the machine translation model from ancient Chinese to modern Chinese, we complete the translation of ancient agricultural books and build a parallel corpus of ancient agricultural books to assist manual tagging and the construction of knowledge graph of ancient agricultural books. Then, this paper systematically combs and organizes the unstructured knowledge of ancient agricultural books through data mining techniques such as knowledge tagging, entity extraction, relationship extraction and so on. Finally, the knowledge graph constructed is not only the reproduction of the knowledge of ancient agricultural books, but also a tool for in-depth exploration and application of this knowledge. The results reveal the effectiveness of the proposed methods in promoting the understanding, application and dissemination of ancient agricultural book knowledge, as well as the importance of interdisciplinary methods in the study of ancient agricultural books.

[收稿日期] 2023-08-14

[基金项目] 国家社会科学基金重大项目"先秦诸子典籍知识库建设及词典编纂"(22&ZD262)

[作者简介] 吴梦成(1997-),男,南京农业大学博士研究生,研究方向为数字人文与机器翻译;

王东波(1981-),男,南京农业大学信息管理学院教授,研究方向为古籍智能信息处理;

黄水清(1964-),男,南京农业大学信息管理学院教授,研究方向为古籍知识组织与文本挖掘。

Key words: ancient agricultural books; machine translation; knowledge labeling; knowledge graph

古农书积累了中国数千年的农业智慧，包括种植、养殖等领域的实践技术和方法，对现代农业生产和科研具有不可估量的价值。然而，这些知识大多散落于非结构化文本之中，其内容的复杂性和语言的深奥性使得这些知识资源的挖掘与利用极具挑战。近年来，得益于人工智能技术的突飞猛进，为解读和利用这些传统知识提供了新的契机。

古农书的翻译不仅仅是语言层面的转换，更涉及对其深厚的农业知识与文化内涵的深入挖掘，这要求译者具备严谨的学术态度与深厚的语言素养。石声汉[①]先生所著的《氾胜之书今释》和《齐民要术概论》的英译[②]，为我国农学史研究奠定了坚实的基础。此外，众多学者在古农书翻译研究领域亦做出了卓越贡献。部分学者主要侧重于典籍译介的分析与梳理，但实际的跨学科翻译实践尚显不足。例如，李海军系统梳理了18世纪以来《农政全书》的英译研究[③]，而王惠琼等人则对20世纪前部分农业科技典籍译介进行了时序性整理[④]。沈思芹等人深入分析了中国古代土壤辨识、桑蚕种养、水土治理等史料的译介情况[⑤]。同时，也有学者聚焦于古农书翻译的科技术语层面，但未能全面覆盖古农书的整体内容。例如，王烟朦探讨了中国古代文化科技术语的翻译方法[⑥]，而袁慧等人则基于译者目的论探讨了科技术语的翻译策略[⑦]。在翻译策略与方法层面，王翠提出中国农学典籍翻译应以研究型翻译为主导[⑧]。徐玉凤从数字英译的视角，归纳了同义对应、同数相对和比例转换三种英译法[⑨]。闫畅等人则强调了中国农业典籍英译中跨文化和跨学科混合翻译模式的重要性[⑩]。然而，尽管学界提出了众多重要的翻译方法，但鲜有学者从信息技术科学的视角对古农书翻译进行探索。

随着信息技术的迅猛进步，古农书的知识组织与挖掘逐渐受到广大学者的关注。在知识组织方面，葛小寒通过分析《树艺篇》与《圃史》的发展历程，揭示了农书知识从获取到整理再到编纂的完整过程[⑪]。徐晨飞等人则以《方志物产》为例，详细阐述了从数字化、数据化、知识化到平台化这四个阶段如何构建知识库并进行深度应用[⑫]。马伟华对《神隐》中的农业知识进行了深入分析，总结出其在农作物栽培、畜牧兽医以及养蚕技术方面的显著成就[⑬]。在知识挖掘方面，陈朝余深入剖析了《三农纪》中的植物保护知识，涵盖了植物防冻法、消雹法等多个方面[⑭]。杜新豪以《便民图纂》中的“耕获类”和“树艺类”为例，揭示了这些农学知识的独特原创性和地域价值[⑮]。李伟霞对《四时纂要》中的兽医知识来源与农医专著中的

① 张保国、周鹤：《石声汉的农学典籍译介模式及其启示》，《解放军外国语学院学报》2022年第5期。

② 孔令翠、曾洁：《石声汉在农学遗产整理与翻译方面的贡献》，《农业考古》2020年第3期。

③ 李海军：《18世纪以来〈农政全书〉在英语世界译介与传播简论》，《燕山大学学报(哲学社会科学版)》2017年第6期。

④ 王惠琼、孔令翠：《20世纪前海外中国农业科技典籍译介研究》，《外国语文》2022年第3期。

⑤ 沈思芹、钱宗武：《〈尚书·禹贡〉所载中国古代农业知识的译介及其西传》，《中国农史》2020年第3期。

⑥ 王烟朦：《基于〈天工开物〉的中国古代文化类科技术语英译方法探究》，《中国翻译》2022年第2期。

⑦ 袁慧、冯炜：《基于目的顺应论的农学典籍〈齐民要术〉科技术语翻译研究》，《长春理工大学学报（社会科学版）》2023年第6期。

⑧ 王翠：《论新时代中国农学典籍的翻译与传播》，《南京工程学院学报(社会科学版)》2019年第4期。

⑨ 徐玉凤：《知识翻译学视域下〈齐民要术〉的数字英译》，《当代外语研究》2023年第6期。

⑩ 闫畅、王银泉：《中国农业典籍英译研究:现状、问题与对策(2009—2018)》，《燕山大学学报(哲学社会科学版)》2019年第3期。

⑪ 葛小寒：《从〈树艺篇〉到〈汝南圃史〉——明代农书生产过程的个案研究》，《自然科学史研究》2020年第1期。

⑫ 徐晨飞、包平：《面向农史领域的数字人文研究基础设施建设研究——以方志物产知识库构建为引》，《中国农史》2019年第6期。

⑬ 马伟华：《朱权〈神隐〉中的农业知识探析》，《科学与管理》2015年第1期。

⑭ 陈朝余：《〈三农纪〉中的植物保护知识》，《农业考古》2000年第1期。

⑮ 杜新豪：《〈便民图纂〉中的农学知识及其价值》，《古今农业》2019年第4期。

各类兽医知识的比重进行了深入挖掘[①]。胡程立以《救荒本草》中的可食用野生植物知识为例，探讨了知识生产与社会实践的相互作用[②]。尽管已有研究覆盖了古农书知识的多个方面，但将现代计算机技术应用于古农书的知识组织与挖掘仍是一个亟待开发的领域。本研究选择七部经典古农书为研究对象，采用知识标注、机器翻译和知识抽取等现代技术手段，构建了古农书的古现平行语料库及知识图谱，并深入探讨了其在知识应用中的巨大潜力。

一、数据来源与研究方法的选择

（一）数据源

在深入探讨中国古农书与机器翻译、知识图谱技术的融合过程中，本研究对数据源的选取进行了严格的考量，力求确保数据源的代表性、准确性和多样性。具体而言，本文从三个核心维度出发，精心挑选适合本研究的中国古农书数据源。首先，在历史代表性与影响力方面，优先选择那些在中国农业发展史上具有重要地位和广泛影响的农书。其次，在版本的权威性和准确性方面，本文特别注重选择由知名专家或学术机构校注的版本。最后，在内容广泛性和多样性方面，本文选取的农书涵盖了广泛的农业主题，如农业技术、作物栽培等多个方面。

以《齐民要术》为例，这部北魏时期的农学巨著，不仅系统总结了当时黄河中下游地区的农业生产技术和经验，而且在中国农业史上具有里程碑式的地位。该书详细阐述了“天时、地利、人和”在农业生产中的重要性，并深入探讨了季节、气候与土壤和作物之间的复杂关系。

基于上述标准，本研究最终精选了《齐民要术》在内的七部具有重要历史和文化价值的农书，为机器翻译模型的微调提供了优质的语料资源，同时也为古农书知识图谱的构建提供了宝贵的数据支持。这些农书的详细信息，如成书时间、作者、版本等，已在表1中详细列出以供参考。

表1　七部古农书语料说明

序号	农书名称	成书时间	作者	包含译文	版本
1	氾胜之书	西汉	氾胜之	否	《氾胜之书今释》本
2	齐民要术	北魏	贾思勰	否	《钦定四库全书》本
3	陈旉农书	宋代	陈旉	否	《永乐大典》本
4	农桑辑要	元代	司农司	是	《钦定四库全书》本
5	王祯农书	元代	王祯	否	《万有文库》本
6	天工开物	明代	宋应星	是	《喜咏轩丛书》本
7	农政全书	明代	徐光启	否	《万有文库》本

此外，为了提升机器翻译模型的训练效果，本文还选用了《二十四史全译》一书作为基础训练语料。该书由两百多名专家历时13年编辑而成，内容涵盖中国历史上的政治、经济、军事、文化、艺术等多个方面[③]。其内容的全面性和深入性，不仅为本研究提供了丰富的历史和文化背景知识，还有助于提升机器翻译模型对中国古代文献的理解和翻译能力。

（二）数据预处理

本研究使用的《二十四史全译》由于未数字化，因此首先采用了光学字符识别（OCR）技术，将其转化为数字文本。鉴于汉字结构的复杂性，OCR识别过程中难免存在误差。为确保古文和译文在句子层面

① 李伟霞：《月令体农书中兽医知识书写特点探析 以〈四时纂要〉为例》，《科学文化评论》2021年第4期。

② 胡程立：《明代农书〈救荒本草〉的作者身份与知识生成》，《出版科学》2023年第6期。

③ 华嘉：《中华文化建设的干城——祝贺〈二十四史全译〉出版》，《民主》2005年第4期。

能够准确对齐，研究采用了Aligner对齐工具进行辅助。在获得初步的OCR识别结果后，研究团队对识别错误的汉字进行了仔细的人工修正，最终获得约100万组平行语料。为提高语料的质量与有效性，本研究进一步使用了相似度检测方法对语料进行筛选。在古汉语专家指导下，我们筛选出古代汉语与现代汉语相似度在0.85~0.98之间的高质量平行语料，共计约30万组。

此外，本研究所使用的七部古农书电子版均通过网络爬虫技术获得，在爬取过程中，我们特别关注了《农桑辑要》与《天工开物》这两部包含原文和译文的农书，实现了原文与译文的同步抓取，而其他古农书则重点抓取其原文部分。值得注意的是，所有农书的注解、注释部分均不在本研究范围之内，故不纳入数据源。

《农桑辑要》与《天工开物》两部农书是微调翻译模型的主要平行语料。在获取数据后，再次使用Aligner对这两部农书进行句对齐处理。通过这一步骤，将原文和译文中较长的句子进行切分，既增加了平行语料数量，又避免了过长句子对模型训练的干扰；同时，结合人工对齐和校正等步骤来提升语料质量，从而提升古农书机器翻译模型的性能。最终使用Python自编程序将语料转化成JSON格式，用于后续翻译任务。《陈旉农书》《氾胜之书》《农政全书》《齐民要术》《王祯农书》这五部包含原文的古农书，也经过了数据预处理和格式化，详见表2。

表2　　农书语料格式

语料类型	语料内容样例
训练语料	{"ancient": "后与秦战，为秦所获，立十四年而死。", "modern": "后来与秦国作战，被秦军捉住，在位十四年而死。"}
微调语料	{"ancient": "崔寔曰：三月可种粳稻。稻，美田欲稀；薄田欲稠。", "modern": "崔寔说：三月可以种粳稻。肥沃的田地，稻子应该种得稀疏一些；贫瘠的田地，稻子应该种得密一些。"}
预测语料	{" ancient ": "豆花憎见日，见日则黄烂而根焦也。", " modern " "}

（三）研究方法与思路

在中国古农书的翻译和知识组织研究领域，传统方法与基于信息技术的方法各有特点。传统翻译方法主要依赖专家学者的逐字逐句翻译和解释，这种方法需要深厚的语言和文化背景知识，但往往效率较低且难以应对大规模古文文本。相比之下，基于信息技术的机器翻译方法，结合了先进的神经网络和预训练模型，不仅加快了翻译过程，还保持了相对较高的准确度。此外，这种方法能够灵活地适应不同文本和语言风格，从而增强了研究的灵活性和广泛性。在知识组织方面，传统方法依赖于专家的手动标注和知识整理，这通常需要花费大量时间并需要深厚的专业知识。然而，利用自动化的知识抽取和智能数据分析技术，可以从大量文本中快速且高效地提取结构化知识[①]。这不仅提高了知识组织的效率，还能发现新的知识链接和模式，为研究提供更深入的洞见。

因此，本研究从信息技术角度出发，采用以下技术路线（图1）以完成中国古农书翻译与知识组织工作：

1. 中国古农书平行语料库的构建

本研究首先聚焦于《二十四史》平行语料，通过对比分析法深入评估神经机器翻译模型与多种基于古籍预训练语言模型的翻译性能，基于这一评估结果确定了最优的翻译模型。随后使用《农桑辑要》与《天工开物》两部农书的平行语料对最优翻译模型进行微调。最终运用微调后的模型对《农政全书》等其余五部农书进行翻译，从而完成中国古农书平行语料库的构建。

2. 中国古农书知识图谱的构建

在成功构建中国古农书平行语料库基础上，对七部农书中实体与关系进行了细致的人工标注，获得了细粒度、结构化的农书数据。接着，利用传统序列标注模型和基于预训练语言模型的方法训练知识抽取模型，选择最优模型实现对七部农书结构化知识的抽取，并结合人工校对确保知识的准确性。最终以

① 吴梦成、林立涛、齐月等：《数字人文视域下先秦典籍植物知识挖掘与组织研究》，《图书情报工作》2023年第12期。

这七部农书的知识及其部分外部特征构建知识图谱并结合Neo4j实现可视化，进行知识检索功能和知识应用的探讨。

本文采用的基于信息技术的方法在处理大规模古代文本方面展现出明显的优势，特别是在效率、范围和深度上超越了传统方法，使其不仅适用于古农书的分析，也为类似的历史文献研究提供了新的研究视角和研究方法（图1）。

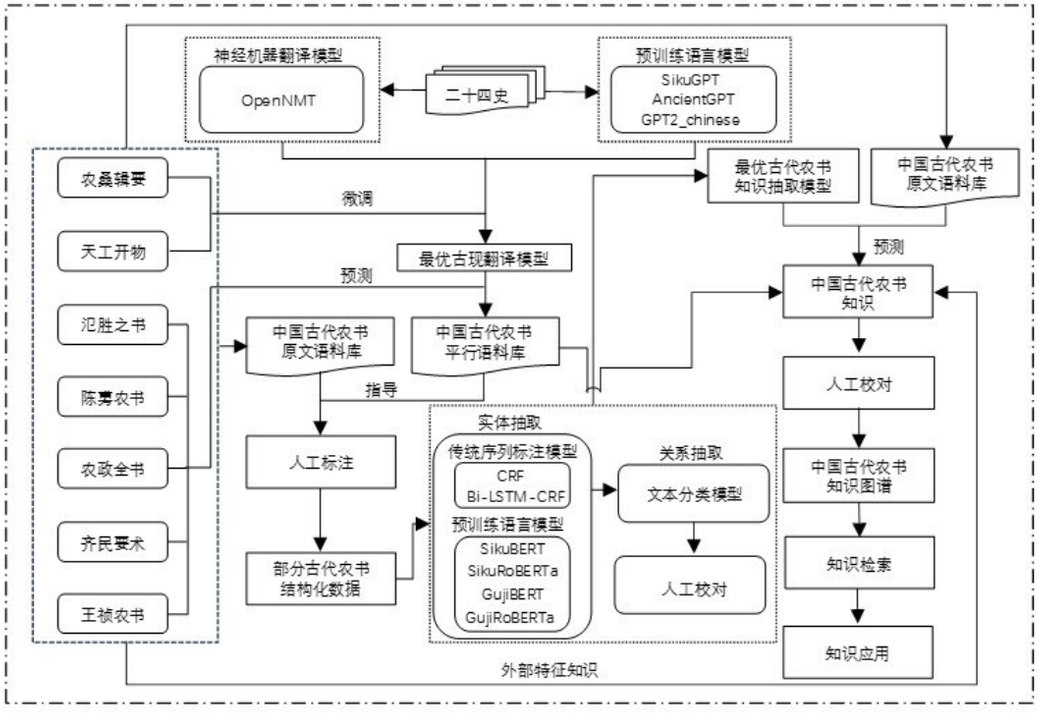

图1 面向古农书的机器翻译与知识图谱构建技术路线图

二、古农书的现代汉语翻译

（一）翻译模型选择

为了提高古农书翻译模型的质量，本研究采取比较分析的方法，即通过对比神经机器翻译模型与多个基于预训练语言模型的翻译模型，确定最优古现翻译模型。这一过程旨在比较不同模型的翻译效果，选择最优模型实现古农书译文的预测，并据此构建中国古农书平行语料库。

在选择最佳翻译模型的过程中，本研究评估了五种主要的模型候选，分别是OpenNMT[①]、SikuGPT[②]、GujiGPT[③]、Ancientgpt[④]和GPT2_chinese[⑤]。其中，OpenNMT是一个开源的神经机器翻译框架，以其强大的功能适合各种规模和复杂度的翻译任务。SikuGPT和GujiGPT是南京农业大学提出的分别基于《四库全书》语料和殆之阁语料开发的生成式语言模型，通过因果语言模型（Causal Language Modeling，CLM）训练而成。AncientGPT，基于GPT-2开发，是一种专注于古文生成的预训练语言模型，利用殆知阁的

① OpenNMT官网地址：https://opennmt.net/。

② SikuGPT模型地址：https://huggingface.co/LC748NLP/SikuGPT2-translation。

③ GujiGPT模型地址：https://huggingface.co/hsc748NLP/GujiGPT_fan。

④ AncientGPT模型地址：https://huggingface.co/Jihuai/bert-ancient-chinese。

⑤ GPT2_chinese模型地址：https://huggingface.co/uer/gpt2-chinese-cluecorpussmall。

古文简体字语料库进行补充标点和训练。GPT2_Chinese则是基于15GB的中文语料库训练的中文版GPT-2模型。

（二）评价指标

在训练面向古农书的古现翻译模型时，使用BLEU[①]（Bilingual Evaluation Understudy）指标和CHRF[②]（Character n-gram F-score）指标作为综合量化的评估方法，帮助衡量模型的翻译准确性和流畅度。BLEU是一种用于评估机器翻译质量的指标，它主要用于衡量机器生成的译文与参考译文之间的相似度。具体而言，BLEU值越高代表翻译模型性能越好。CHRF也是一种评估机器翻译系统质量的指标。与BLEU最大的区别在于CHRF以字为单位对翻译质量进行评估，而BLEU是词级别的翻译质量评估方法。

（三）参数设置

本实验所需的计算机配置如下：操作系统为CentOS 3.10.0，CPU为4颗Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz，内存大小256G；GPU为6块NVIDIA Tesla P40，显存大小24G。实验参数配置具体如表3所示。

表3　超参数配置信息

实验超参数	参数定义	设定值
max_sequence	最大输入长度	512
learning_rate	学习率	1e-5
max_len	最大生成长度	256
batch_size	批大小	12
max_epochs	最大迭代次数	30

（四）古农书平行语料库构建

通过上述实验，发现使用古农书平行语料微调后的GujiGPT翻译模型，实际翻译性能最高，BLEU值达到32.73，CHRF值达到29.19。具体各模型翻译效果的评测结果见表4。

表4　古农书在各模型翻译性能评价结果

模型名称	BLEU	CHRF
OpenNMT	25.36	23.72
SikuGPT	28.13	25.35
GujiGPT	32.73	29.19
AncientGPT	27.98	25.19
GPT2_chinese	28.71	25.57

因此本研究利用该模型实现了《农政全书》等五部农书的翻译，部分最终翻译结果见表5。结合《天工开物》与《农桑辑要》两部书的平行语料，本研究共获得七部古农书的平行语料。据此，在垂直领域上成功构建中国古农书平行语料库。

① Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Trans - lation: Research Papers, pages 186 – 191, Brussels, Belgium. Association for Computational Linguistics.

② Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392 – 395, Lisbon, Portugal. Association for Computational Linguistics.

表5 五部农书的现代汉语翻译结果示例

农书名称	原文	译文
《陈旉农书》	古人种桑育蚕，莫不有法。	古人种桑树养蚕，都有规定。
《氾胜之书》	种伤湿郁热则生虫也。	种植受到湿郁热害，便会生虫。
《农政全书》	尝闻古耕者用耒耜，以二耜为耦而耕。	曾听说古代耕田的人用耒和犁，是用两个犁来作为一个专门农具。
《齐民要术》	神农之时，天雨粟，神农遂耕而种之。	神农氏时，天上降下粟米，神农于是翻耕土地。
《王祯农书》	胡麻，即今之脂麻是也。汉时张骞得其种于胡地，故目之曰胡麻。	胡麻，就是今天的油脂麻。北方人张骞得到种子在胡地，所以把它叫做胡麻。

三、古农书的知识组织过程

（一）古农书知识抽取

本研究主要通过Inception工具来完成古农书中知识的标注。Inception是一个文本标注环境，可用于对书面文本执行各种标注任务，标注内容通常是关于语言或机器学习的问题，并且该工具提供推荐系统，有利于更快、更便捷地创建标注对象。同时在人工标注过程中，将古农书的译文作为主要参照以辅助确定标注内容的准确性。

1. 古农书实体标注规范

本研究对古农书中的植物、动物、农具、技术、土地、时间、气候、灾害、人名、地名、农事活动这十一种实体进行了细致地标注。在此标注过程中，植物类实体不仅包含了古代农作物如稻谷、小麦等，还涵盖了自然生态中的植物，如“松”“竹”等，以反映当时的生态环境。动物类实体主要囊括了鸡、猪、牛等家畜，以及狐狸、兔子等野生动物，体现了古代农耕与狩猎的生活方式。农具实体则涵盖了犁、耙等耕作工具，技术实体包含了灌溉、施肥等农业技术，凸显了古代农业生产的关键要素。土地实体考虑了不同类型的田地、土地类型，时间实体覆盖了季节、节气等时间概念，气候实体涵盖了干旱、雨水等气候特征，而灾害实体包含了洪水、虫灾等自然灾害以及瘟疫等人类健康灾害。人名实体涵盖了在农书中有显著贡献的历史人物，地名实体反映了重要的地理位置，农事活动实体则包含了耕种、收割等具体农业活动。这一精细的实体标注工作为本研究构建了一个更加丰富、精确的古农书知识体系，也为构建古农书知识图谱提供了坚实的基础。

2. 实体数据统计与分析

为构建高效的古农书实体抽取模型，本研究邀请6名具有古文标注经验的研究生对七部农书中的十一类实体进行人工标注。由于古籍数量较多，语料规模较大，因此标注过程中只对每部农书的20%左右数据进行标注。同时为保证标注质量，标注人员被分为三组，每组两人，在农书译文的指导下，对每部农书逐句进行标注。第一轮标注完成后，每组同学将各自标注内容进行交换，对对方的标注结果进行检查和校正，最大可能保证最终农书实体标注的质量。最终各实体标注结果如表6所示。

表6 古农书各实体人工标注统计结果

序号	实体名称	实体示例	实体总数	不重复实体数量
1	植物	粟、杨、荆、李	1150	368
2	时间	唐、宋、春、夏、秋、冬	331	98
3	动物	牛、羊、马、犬	308	51
4	人名	神农、李悝、尧、舜、禹	229	103

续表6

5	农具	耜、耒、犁、锄、镰	189	37
6	土地	腊田、脯田、黑垆土	158	63
7	技术	踏粪、冬藏雪汁	160	78
8	气候	雨、雪、风、寒	147	40
9	地名	江陵、齐鲁、渭川	159	77
10	灾害	水、旱、蝗虫、潦	75	20
11	农事活动	杀、耕、种	273	21

3.古农书实体自动抽取

古农书是属于中国古代典籍的一个重要分支领域。因此，为提升该领域的实体识别效果，在模型选取方面选择经过古文预训练过的模型如SikuBERT、SikuRoBERTa、GujiBERT、GujiRoBERTa等对古农书的特定实体进行抽取。实验具体结果如表7所示。

表7　古农书各模型实体抽取结果

模型名称	P	R	F1
CRF	0.74	0.70	0.72
Bi-LSTM-CRF	0.75	0.71	0.73
SikuBERT	0.77	0.84	0.80
SikuRoBERTa	0.74	0.81	0.77
GujiBERT	0.78	0.85	0.81
GujiRoBERTa	0.73	0.82	0.77

由表7可知，本实验中GujiBERT模型对农书中的11类实体识别综合效果最优，达到81%。因此研究选用该模型，对七部农书中剩余80%左右未标注语料，进行实体识别并结合人工校对的方法将未识别出和识别错误的实体进行补充和更正，最终七部农书中的实体统计数据见表8。可见，在十类实体当中，古农书最常出现的实体是植物实体，总数高达35271个，不重复植物实体为2363个。

表8　古农书各实体最终统计结果

序号	实体名称	实体总数	不重复实体数
1	植物	35271	2363
2	时间	7828	394
3	动物	7657	200
4	人名	4075	867
5	农具	6503	200
6	土地	943	114
7	技术	3569	98
8	气候	10640	141
9	地名	8923	717
10	灾害	1180	22
11	农事活动	8442	45

4.古农书关系自动抽取

不同于现代汉语文本，古农书文本中实体关系通常不会直接在原文中明确表示，而是需要通过上下文和语境来人工推断。因此，基于规则的方法与通用的关系抽取模型并不能适用于古文领域的关系抽取。在当前任务场景中，关系抽取其实是指从给定古农书文本中，识别和提取出特定实体之间的关系。

因此，在本研究中为厘清各类实体之间的确切关系，采用文本分类的方法实现这个任务。具体地，将每个句子作为输入，通过训练模型来识别古农书中的关系，并对这些关系进行分类。模型可以学习关系之间的共同模式和上下文信息，并在给定句子时预测出最有可能的关系类型。

为有效构建训练语料，本研究将含有实体的5000余古文句进行整理。利用给定实体关系集人工标注其中1500句的实体关系。输入语料采用（古文句+空格+实体集）的结构以便于模型习得实体在古文句中的语义关系。部分输入语料如表9所示。

表9　古农书关系抽取输入语料格式

序号	输入语料	关系标签
1	高田早稻，自种至收，不过五六月。（高田，早稻）	土地–影响–植物
2	二月种粟。（二月，粟）	时间–适合–植物
3	季夏之月，利以杀草，可以粪田畴，可以美土疆。（季夏，杀草）	时间–适合–农事活动

经过5轮模型训练，给定实体集的古农书实体关系分类结果调和平均值高达近70%，在训练完成后，利用最终模型对所有含有实体的古文句进行关系抽取。对抽取结果进行多轮人工校对，确定每句的实体关系，最终主要实体关系统计结果见表10。

表10　古农书关系抽取结果

序号	实体关系名称	实体关系数量
1	气候–影响–植物	217
2	灾害–影响–植物	46
3	土地–影响–植物	13
4	时间–出现–气候	375
5	时间–适合–农事活动	338
6	时间–适合–植物	1,116
7	农事活动–用于 –植物	1,250
8	农事活动–应用–技术	72
9	农事活动–使用–农具	137
10	技术–用于–植物	224
11	技术–防治–灾害	40

整体而言，“时间–适合–植物”和“农事活动–用于–植物”在七部农书中的实体关系更加普遍。而“灾害–影响–植物”“土地–影响–植物”“技术–防治–灾害”等实体关系较少。

（二）知识表示

本研究中标注的古农书知识主要分为两部分，一部分是实体知识包括农书中的植物、动物、农具、技术、土地、时间、气候、灾害、人名、地名、农事活动十一种实体。另一部分是实体关系包括“气候–影响–植物”“灾害–影响–植物”“土地–影响–植物”“时间–出现–气候”“时间–适合–植物”“时间–适合–农事活动”“农事活动–用于–植物”“农事活动–应用–技术”“农事活动–使用–农具”“技术–用于–植物”“技术–防治–灾害”十一种关系。通过三元组的形式将上述知识进行整理和表示，具体如图2所示。

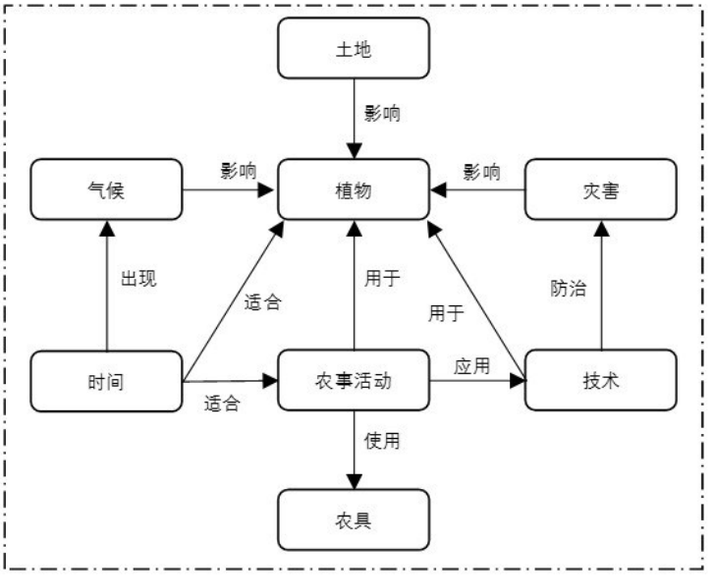

图2 古农书实体关系知识表示

(三)知识融合

知识融合是指整合、合并和验证来自不同数据源的知识,以构建一个一致、完整、准确的知识库或知识图谱。在知识融合的过程中,不同的数据和知识被结合在一起,以获得更全面、准确的知识体系。在构建古农书知识图谱的过程中,需要将从农书中抽取的实体和实体关系等知识进行融合,具体是将各信息源中重复、不一致与含有冲突的知识进行分析、比较、验证和整合。譬如,在实体层面,从《氾胜之书》和《农桑辑要》中提取的表示气候变化的实体词包括“冻解”和“冻释”。这两个词在语义上具有相似性,且都与时间词“春”有关。因此,在知识融合过程中,需要对这两个相似的实体进行比较、验证和整合,确保只保留一个准确、一致的实体表示。在实体关系层面,《氾胜之书》中存在表示时间与气候关系的实体关系“春-影响-冻解”,而在《农桑辑要》中也存在表示时间与气候关系的实体关系“春-影响-冻释”。因此在知识融合过程中,需要对这些实体关系进行比较和融合。此外,除了在农书内部进行知识融合,还需要将农书的名称、作者以及农书的译文等外部知识一并纳入融合,以形成最终农书知识图谱构建所需的完整知识体系。

(四)知识存储

知识存储是将整合、验证和融合后的知识以结构化的方式进行组织和存储的过程。在古农书知识图谱的构建过程中,知识图谱作为一种强大的工具,可以有效地存储和表示各种实体及其之间的关系,提供灵活而直观的知识管理方式。通过知识图谱的存储,古农书中的各种实体,如植物、动物、农具、技术等,可以作为节点进行表示。同时,实体之间的关系,如气候-影响-植物、时间-适合-植物等,可以用边连接不同的节点,形成一个具有丰富关联性的知识网络。知识图谱作为一种存储古农书知识的方式,能够以结构化、语义化的形式呈现实体与实体的关系。通过建立关联性强且可扩展的知识网络,为共享、研究与应用古农书的知识提供了有力支持。

四、古农书的知识检索与应用分析

(一)知识检索

尽管通过数据挖掘等技术实现了对中国古农书知识的有序组织和高效存储,然而本研究的最终目

标并非仅限于这些知识的收集与整理。实现知识的共享与传承，让这些古农书知识“活”起来，才是数字人文研究真正的价值所在。为此，综合考虑了关系型数据检索的复杂性、灵活性和扩展性等多个方面，选择使用Neo4j工具，以三元组的形式对农书知识进行了系统梳理与存储。通过图数据库中的三元组，将农书知识的不同要素相互连接，形成一个相对完整且高度关联的知识网络。在知识网络中用户可以通过实体或者实体关系对目标知识进行检索并对检索结果进行分析。

本文从古农书中随机抽取出时间实体“春”，使用Cypher语言命令“MATCH (n {name: '春'}) - [r] - (related) RETURN n, r, related;”对与“春”可能有关的知识进行检索。从图3可以看出，“春”这个时间段适合养蚕，适合种植豆、大麦、山药等农作物，经常会发生的灾害有蝗虫灾害和洪涝灾害，在气候上通常是多风、多雨，并且在这个时间段除草、耕地是比较常见的农事活动。

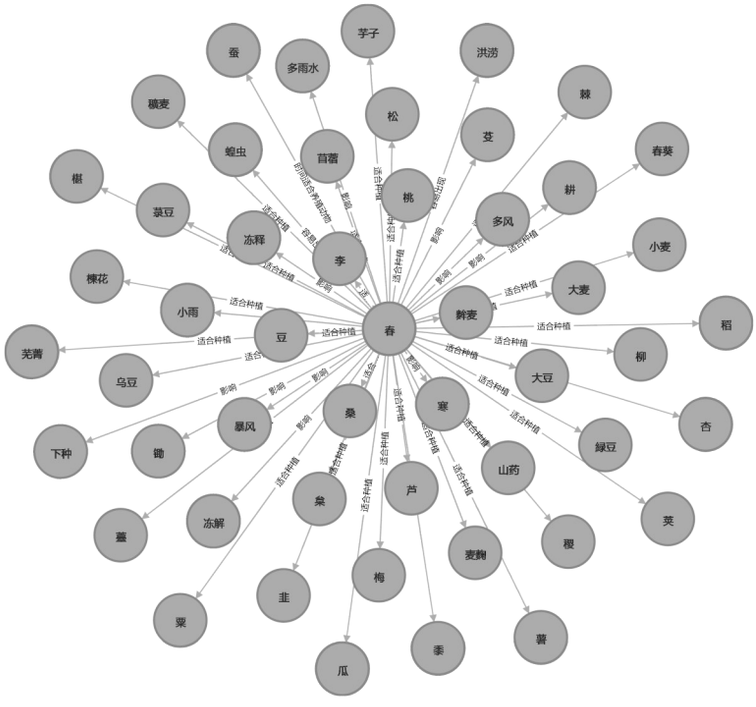

图3 古农书图数据库实体检索图例

同时，为了确保检索结果的完整性和可靠性，本研究在图数据库中不仅引入了农书中的细粒度实体知识，还以实体与原文的关联为枢纽融合了农书名称、农书作者等题录数据。一方面有利于用户更加全面地了解农书知识的来源和背景，另一方面也为研究者提供了更多维度的查询和分析手段。图4展示了使用Cypher语言查询古农书中与农事活动“刈”关联的部分知识。可见“刈”可用于植物“穄”和“黍”，并且该实体及关系见于元代司农司撰写的《农桑辑要》一书。

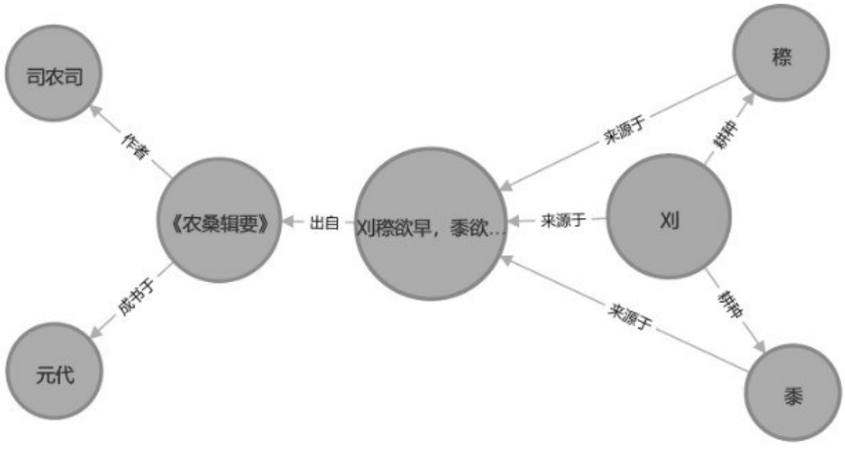

图4 古农书图数据库知识检索图例

此外，在知识网络中还引入了古农书“译文”节点，将古文原文与其相应的现代文译文进行关联，以建立实体知识与多语言表达之间的桥梁，为难以理解的实体提供辅助理解的途径，促进知识的传播与共享。通过这样的关联机制，用户在查询农书知识图谱时，不仅可以获取实体的内容与关系，还能直接查看古文原文和对应的现代文翻译。具体如图5所示，在“三月适合种植粳稻”这组实体关系中，“粳稻”植物实体来源于“南方大麦有既刈之后乃种迟生粳稻者。”古文原文并与其译文“南方的大麦收割后才会种植迟生的粳稻。”相关联。这种直接关联的方式为用户提供了更加全面和深入了解农书知识的途径。对于那些难以理解的古农书术语或概念，用户可以通过查阅其原文和译文，追溯知识的来源和含义，进一步增进对农书知识的理解。此外，译文节点的引入还为知识的传播与推广提供了有益支持。通过译文节点，农书知识可以现代文的形式传播，跨越语言和时间的限制，让更多人受益于古代智慧。

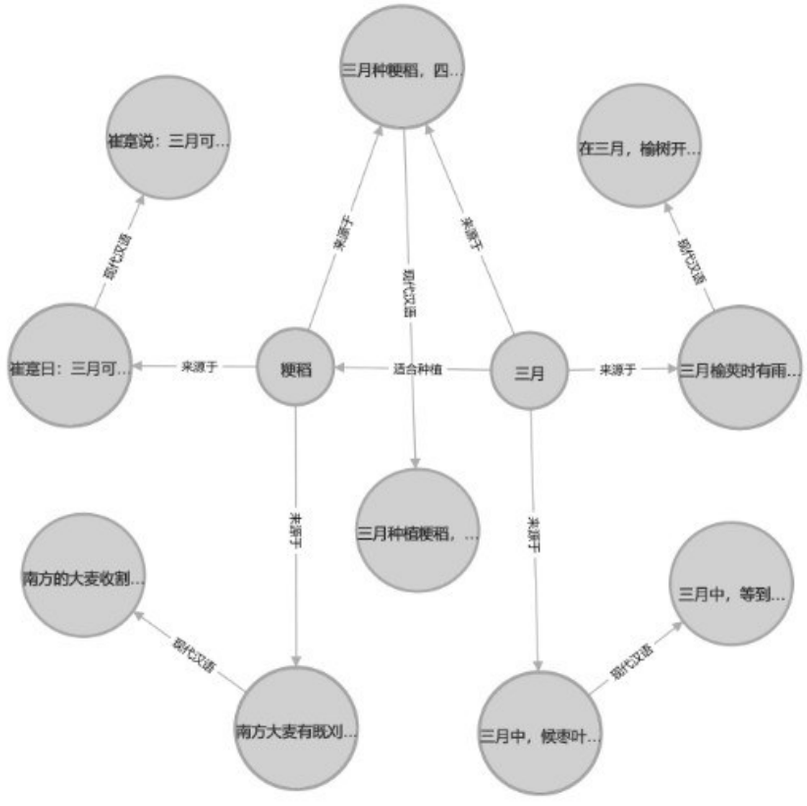

图5 古农书图数据库译文节点图例

可见通过该图数据库中的查询功能，用户可以轻松找到与自己研究相关的农书实体、关系以及其对应原文、译文信息。同时，图数据库具有较强的扩展性，因此关于农书方面新的研究成果可以与现有知识相结合，从而进一步丰富和完善古农书知识体系。

（二）应用分析

古农书知识图谱中蕴含丰富的隐性农书知识，结合统计分析、时序分析等数据分析技术可以深度挖掘古农书中的动态变化和复杂关系。

从静态视角看，古农书知识图谱为农学和农业研究者带来了一个强大的统计分析工具，能够深入研究古农书的知识体系，挖掘其中隐含的重要知识。古农书知识图谱可以统计出古农书中植物、动物、农具等实体的出现频次信息。结果显示，出现频次最高的动物是蚕，而不是人们通常所认为的牛，蚕共计出现了1527次。同时出现频次最高的植物是桑，频数为1308次。这个发现进一步印证了桑与蚕在古农书中紧密的关系。桑树作为蚕的主要食物来源，与蚕的饲养密切相关，而蚕作为丝绸生产的关键，从蚕

茧中提取的丝线构成了珍贵的纤维材料。这种相互依存的生态链条在古农书的记载中得到了充分的体现，彰显了当时人们对于丝绸产业的深刻认识和应用。因此，古农书知识图谱作为统计分析工具不仅可以使我们能够更全面理解农书中的知识关联，还有利于揭示古代农业知识的丰富内涵，甚至挖掘出新的知识。

从动态视角看，古农书知识图谱可以从时间维度揭示不同历史时期农书中某类农具、技术等发展的历史轨迹及趋势。通过分析古农书知识图谱，可以统计得到西汉、北魏、宋元明三个时期农具的具体使用情况。由于不同时期农书语料规模存在差异，因此采用农具在相应时期农书中出现农具总数的比例，来衡量不同历史时期农具的使用热度。研究结果显示，在西汉、北魏、宋元明三个时期，农具种类数量持续增加，从9种增至70种，再扩展至200种。特别是“锄”，在各时期均是使用最广泛的农具，其使用热度分别约为63.3%、12.4%、9.8%，体现了“锄”在中国农业生产中的重要性和普及程度。另外，“锯”“镰”“剗”这三种农具在西汉时期也有较高的使用热度，约6.7%。而在北魏时期，这三者的热度分别下降至约1%、0.7%、0.3%。到了宋元明时期，其热度分别提升至约为1.9%、1.1%、0.4%。这反映出随着农具种类的增加，更多高效的农具被发明，农业生产方式也随之发生了变化。比如，“斧”被用来砍劈木材，“杵”用于研磨谷物，“犁”用于松动土壤等，这些新型农具在历史上一直保持着较高的使用热度。此外“磨”作为粮食加工的主要工具，在北魏时期使用热度仅为2.9%，而到了宋元明时期增至5.3%。这可能意味着相应的粮食加工技术在这一时期得到了显著发展。

这些分析技术的应用不仅让后人更深入地理解古代农业知识，还能够促进现代农业技术的创新。例如，可以结合地理信息系统技术，分析不同农具在特定地区的使用情况，并根据当地的气候条件、土壤类型或作物种植习惯进行更深层次的知识挖掘。这样的发现可以为现代农业提供历史参考，帮助农业工作者选择更适合当地条件的农具和耕作方式。同时，结合关联规则挖掘技术，揭示不同农具之间的使用关系，例如，耕作工具与收割工具的使用频率是否有相关性，或者某种特定农作物的种植是否经常与特定农技的应用同时出现。这种分析可以帮助后人理解古代农业生产的链条，如耕种、施肥、灌溉、收割、储藏等各个环节之间的相互依赖和优化过程。因此，古农书知识图谱不仅为跨学科研究开辟了新视野，也为现代农业实践提供了宝贵的参考。

结　语

本研究致力于将现代智能信息处理技术应用于古农书的翻译与知识图谱构建，旨在识别、整理、组织并存储古农书的农业知识，有效促进这一重要文化遗产的利用和传承。通过训练和优化机器翻译模型，成功构建了古农书平行语料库，实现了古农书的高效翻译。同时，利用知识抽取技术，构建了古农书知识图谱，并通过Neo4j等工具实现了图谱的可视化，使古农书的农业知识更加易于理解、应用和传播。

然而，本研究仍存在一些不足之处。古农书语言和内容的复杂性使得机器翻译的精确度有待进一步提高。此外，知识抽取过程中也可能存在误差和遗漏，需要持续优化和改进。为了更全面地展示古农书知识的内在关联和应用价值，还需进一步丰富和完善知识图谱的构建和分析。未来的研究可以继续改进机器翻译模型，提升古农书翻译的准确性和流畅度，以更好地满足用户需求。同时，探索更先进的知识抽取技术，结合自然语言处理和大语言模型等方法，提高知识的准确提取和自动化整理能力。此外，可以扩充知识图谱内容，融入更多农业知识，并结合新型技术与方法拓展其应用范围和深度，以促进中国古农书知识和文化的传承与传播。

（责任编辑：李良木）