

当古籍遇见青春：AI 如何让《四库全书》活起来

年轻人的事业

古籍数字化,总共需要分为几个步骤?如果是简单操作,扫描古籍,制成电子版即可。网上流传的数字古籍多为此类。当然,不要对质量抱太大希望。“版本不明、图片模糊是常见问题,给学界和使用者造成了极大困扰。”上海古籍出版社数字编辑室副主任鲁秀梅说。尤其令专业人士无法忍受的,是夸张的差错率。鲁秀梅告诉记者,市面上的数字古籍差错率远远超过图书编校差错率不高于万分之一的国家标准,使用者稍有不慎就会“踩坑”。

不过,若要提供可靠且实用的数字古籍,事情就变得复杂起来。扫描只是第一步,接下来要进行数据的校验和加工,在确保准确率的同时,还要研发翻阅、检索、标点、引用乃至翻译等功能,便于学者开展研究。一两本古籍还好说,可如果是卷帙浩繁的《四库全书》呢?

《四库全书》号称“全书”,是因为它基本囊括了我国清代中期前的所有图书。为推动这项浩大工程,清乾隆帝召集360多名官员、学者进行编撰,3800多人抄写,历时19年才告完工。编成之日,乾隆帝命人手抄7部,分藏北京、沈阳、杭

州等地。近代以来,中国饱经忧患,《四库全书》很难得到妥善的保存,多份抄本或毁于战火或流散民间,最终只剩“三部半”。其中,原藏于故宫的文渊阁本是较为完整的一部,1987年由上海古籍出版社在中国大陆影印发行,共收录3000余种书,分装1500多册,总字数约8亿。

将文渊阁《四库全书》数字化,远不止把纸质书搬到线上那样简单。以前,识别、校对等工作常交给外包,如今几乎所有流程都要由出版社完成——这是一家传统出版社能完成的吗?当社里将打造“文渊阁四库全书数据库”的任务交给数字编辑室时,鲁秀梅内心不免忐忑。但她坚信一点:“数字出版是年轻的事业,要由年轻人挑大梁。”

2019年,为建设“汇典”项目,在上海世纪出版集团的支持下,上海古籍出版社“招兵买马”,很快搭建起一支平均年龄32岁的年轻团队。数字编辑室共11名成员,年长的四十岁出头,最小的1999年出生,其“胶原蛋白含量”大约是整个出版社最高的。这也是一支跨学科团队。成员除了来自文献学、古代文学、历史学等对口专业,还有计算机、机械工程、信息工程等理工专业,堪称“文理兼备”。

“驯服”古籍

高典可以说是文理兼备的“本体”。长着一张娃娃脸的他生于1989年,本科就读于计算机专业,硕士阶段转入中国古典文献学。为什么由理转文?他嘿嘿一笑:“从小听家里的话念理科,但我一直对古代文学非常感兴趣,考研的时候想试试,结果考上了。”巧的是,古典文献学逻辑性极强,高典的理科思维反而成了优势,专业转得十分顺畅,没经历什么痛苦。

毕业后高典在北京工作了一段时间,2021年加入上海古籍出版社。他刚开始做编辑,也处理过疑难字,后来被聘为技术岗,成为文渊阁《四库全书》数字化项目的技术负责人。

“几乎是白手起家。”回忆刚接手项目时的情景,高典感叹,“手头只有文渊阁《四库全书》的扫描图,别的什么都没有。”与现代书不同,古代没有标准化的书写和排版格式,不同书籍往往从版式到装帧都大相径庭,再加上批注、表格和插图,导致样貌千差万别。拿现代书的OCR(光学字符识别)技术去识别,不仅视觉呈现上异常凌乱,更会给使用者制造各种麻烦。

“解决这个问题,必须先确定数据格式。”高典说。于是,他带着技术组的伙伴开始攻关。

这时候《四库全书》的优点体现了出来。它用浙江出产的上等开化纸缮写,即便是两百多年后影印的文渊阁本,依然崭新如初。页面布局也很用心:每页8行、每行21个字,如需注解,标以双行小字。“格式规整,字迹和插图清晰,正好给我们练手,为古籍数字化攒经验、树模板。”高典说。这也是上海古籍出版社决定率先将文渊阁《四库全书》数字化的原因之一。

饶是如此,确定数据格式还是花了他们两个多月的时间。经过反复推敲、反复尝试,他们自主研发的OCR工具也有

了雏形,可以顺利地把古籍图片中的文字识别并提取出来。

到了这一步,真正的难题,才刚刚浮现。古籍里存在大量异体字、通假字、避讳字,古籍影印版中也不可避免会有字迹模糊的情况,OCR经常识别错误,这就需要人工校验。更大的挑战来自插图。“古籍的插图很多时候是文字,但它既然以图的形式表示,就不能把它当文字。”高典说。可OCR哪里分得清?于是又要人工判断。据估算,数字编辑室组织识别了20多万个疑难字,插图加表格超过10万页。所幸,数据化工组的三名95后成员,凭借古典文献学的专业功底,起到了把关作用。

2024年初,“文渊阁四库全书数据库”上线,支持全文阅读、检索、引用复制、联机字典、纪年转换等功能。这群年轻人交出了“驯服”古籍的初步成果。



上海古籍出版社数字编辑室团队成员合影。



《四库全书》。

本版均为受访者供图

和“幻觉”作斗争

“在完成‘文渊阁四库全书数据库’的基础上,我们希望能与AI技术结合,为用户提供更优质服务,因此决定打造‘文渊阁四库全书’AI+版。”鲁秀梅说。这意味着除了阅读和检索等基本功能,还要能AI标点、AI标引、AI翻译……这次,轮到另一名“技术大牛”陈诚大展身手了。

陈诚,1997年出生,机械工程硕士,毕业后先在某大厂干了两年多,主攻自动驾驶。大厂的工作强度很大,陈诚觉得没什么业余生活,刚好上海古籍出版社招聘算法工程师,自己又对古文感兴趣,就投了简历,顺利加入数字编辑室。恰逢“文渊阁四库全书”AI+版启动,有AI项目实操经验的他立刻挑起大梁。

“古籍坐拥大量优质古籍资源,为我们提供了丰富的高质量语料,能够喂给大模型,构造数据集。”陈诚介绍。以AI标点为例,他们找来已经出版的书,去掉标点,先让大模型添加标点符号,随后

对照原书一点点校正。经过不断地训练和调试,AI标点渐渐有模有样,正确率可达97%。这已然专业水平了。

真正的难关在文言文翻译。“AI的幻觉太严重了,经常翻译得不知所云。”陈诚说。比如,《阿房宫赋》最后一句:“秦人不暇自哀,而后人哀之;后人哀之而不鉴之,亦使后人而复哀后人也。”大模型怎么都翻译不好。

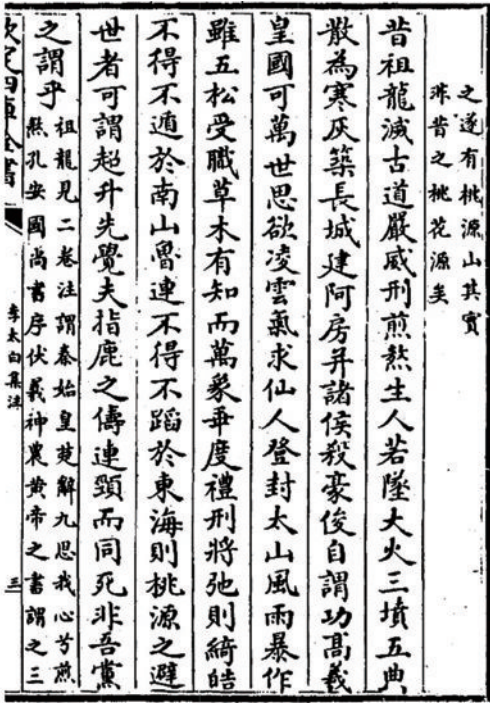
“大模型其实是个‘黑盒子’,我们不知道它能输出结果,但搞不清里面是怎么运作的。”陈诚解释。简单来说,虽然大模型本质是一种算法,它的参数却不是程序员提前设定的,而是其自主学习和调整形成的。由于参数量极为庞大,整个系统表现出的数学复杂性超出了人类能直观理解和推理的范畴。

情况就是这么个情况,阁下将如何应对?

只能下“笨功夫”。数字编辑室的小伙伴们群策群力,用一篇又一篇古文去测试大模型,一字、一词、一句地琢磨它到底哪里犯错、为什么会犯错,确认后再用提示词进行约束和规范。“跟打地鼠一样,出来一个打一个,靠提示词压制幻觉。”高典说。这样靠着几百篇古文的反复测试,翻译的准确率大大提高。

与此同时,高典策划了流式阅读,让使用者能方便地调用AI功能。流式阅读还使古籍文本以适合当代读者阅读习惯的形式呈现,既实用又美观。

“文渊阁四库全书”AI+版毕竟是大工程,尽管有出版社的全力支持,仅靠数字编辑室的力量还是不够,必须借助技术公司及数据加工团队。但外包方不熟悉古典文献,如何让他们“听懂”需求呢?产品经理王子伟发挥了作用。她本硕念的都是中国古典文献专业,专业过硬,求学期间又用过多款古籍数据库,善于站在用户角度看问题。通过她的内外衔接,数字编辑室和外包方得以紧密合作,项目“丝滑”推进。



古籍原文。

AI回答 自动翻译结果

从前秦始皇毁灭古代的道义，严刑峻法折磨百姓，如同陷入烈火之中。“三坟五典”等古籍化为灰烬，他修筑长城、建造阿房宫，吞并诸侯国，杀害豪杰俊才，自认为功绩超过伏羲和皇帝，国家可以万世不衰。他想要凌驾云气之上，寻求仙人长生不死，在泰山举行封禅大典时，风雨突然暴发。虽然五棵松树被任命为官员，传说秦始皇在泰山封禅时，命五棵松树为“五大夫”，草木尚有灵性，但天地万物的秩序已经紊乱，礼仪和刑法也将松弛。于是始皇季、夏侯蒙不得不逃到南山；鲁仲连也不得不投海自尽。那么那些避世于桃花源的人，可以说是超脱尘俗、先知先觉之人了。至于那些指鹿为马之徒，甘愿同流合污而死的，不是我们这一类人吧！

AI回答 自动标点结果

昔祖龍滅古道，嚴威刑，煎熬生人，若墜大火。三墳五典散為寒灰，築長城，建阿房，并諸侯，殺豪俊，自謂功高義重，國可萬世。思欲凌雲氣，求仙人，登封太山，風雨暴作。雖五松受職，草木有知，而萬象乖度。禮刑將弛，則綺結不得不遁於南山，魯連不得不蹈於東海。則桃源之避世者，可謂超升先覺夫！指鹿之偽連預而同死，非吾輩之謂乎？

成果显而易见,挑战接踵而至

青年报记者 唐骋华

本报讯 2025年8月,“文渊阁四库全书”AI+版问世,并现身上海书展,吸引了众多读者前往体验区观看和试用。一名中学生轻轻用指尖在大屏幕上划了一段古文,AI立刻翻译成现代汉语。一名学者凑近屏幕,看着AI自动标点的古籍原文连连点头,编辑递上一张30天试用卡,笑着说:“回家您慢慢研究。”

一同亮相上海书展的还有“典籍整理文献数据库”和“上海文献总库数据库”。前者以下海古籍出版社出版的核心整理本古籍为主,同时收录上海世纪出版集团内、外相关出版机构的优质资源,包括中国古典文学丛书、十三经注疏、商周青铜器铭文暨图像集成系列等。后者收录上海府县旧志丛书、上海市新编地方志书,全面记述上海的历史与现状。加上“文渊阁四库全书数据库”,共计约20亿字,由此构成了“尚古汇典·古籍数字服务平台”的内容部分。

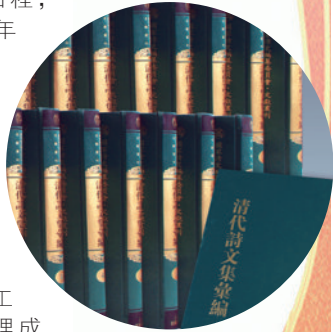
“上海古籍出版社在数字出版市场急缺古籍精品的情下,主动承担起补缺工作。”鲁秀梅如此阐释打造“尚古汇典·古籍数字服务平台”的意义。目前,该平台已在技术上完成古籍OCR个人版、机构版的优化开发,以及AI自动标点、自动标引、自动翻译等功能的适配与优化。值得一提的是,由上海古籍出版社

自主研发的OCR识别系统已在四川大学等高校及科研机构的文献整理项目中得到应用,极大地提升了效率。

正所谓百尺竿头更进一步。陈诚坦言,“文渊阁四库全书”AI+版尚有改进空间,“比如AI幻觉,还远远谈不上根治”。他和伙伴会继续调试,将“幻觉”控制在合理范围内。王子伟则表示,自己在产品开发和运营方面积累了一定经验,有了不少想法,希望在今后的工作中加以实践运用。

而这群年轻人即将迎来新的挑战。鲁秀梅告诉记者,《清代诗文集汇编》数字化项目已提上议事日程,预计于2026年启动。

《清代诗文集汇编》是上海古籍出版社于2010年影印出版的大型清人文集,是清史工程的重要整理成果,收书约4000种,共4亿字。《《清代诗文集汇编》的数字化将是古籍社继‘文渊阁四库全书数据库’后的又一盛举。”鲁秀梅说,“当然工作量也是空前的。”但她有信心,凭借优化后的OCR识别系统和工业化流程,不仅数字化速度和准确率会提升,整体成本也将缩减三分之二。



【转型之路】

从0到1乃至无穷

回想往昔,鲁秀梅颇为感慨:“数字编辑室成立之初,主要工作是承担社里的电子书业务,建造数据库的经验为零。”从2020年到2025年,经过近6年磨砺,这支年轻的团队打造了“尚古汇典·古籍数字服务平台”,建成“典籍整理文献数据库”“上海文献总库数据库”“文渊阁四库全书数据库”三大数据库。这是从0到1的一步,更是从1通向无穷可能的起点。

这也折射出一家传统出版社的转型之路。成立近70年,上海古籍出版社整理出版了众多优质古籍,专业性和权威性有目共睹,是古籍出版领域公认的头部品牌。但时至今日,即便头部也面临汹涌的数字化浪潮。“出版业必须紧跟时代步伐,拥抱数字化变革,研究当代读者、用户的深层次精神文化需求,提供更好的知识服务。”上海古籍出版社副社长兼副总编吴长青说。

在吴长青看来,当年上海古籍出版社影印出版文渊阁《四库全书》,为古籍的校勘和研究提供了

重要参考底本,得到学术界盛赞。今天,出版社的青春力量为文渊阁《四库全书》插上AI的翅膀,使其既能便捷高效地辅助用户阅读与理解专业类古籍文献,也满足了用户对古籍阅读的一般需求。更重要的是,一批后起之秀在这个过程中冒尖。“比如鲁秀梅,早年做纸质书编辑,后来参与数字化项目,带领数字编辑室,为全社储备了数字化人才。”吴长青说。从这个意义上说,“尚古汇典·古籍数字服务平台”其实是上海古籍出版社数字化转型的抓手,将带动全社理念、人才和流程的变革。

华东师范大学中文系教授、古籍研究所所长方笑一则从学术角度肯定了上海古籍出版社的实践。他认为,古籍数字化一方面让珍贵文献突破时空限制,实现文本的永久保存和便捷查阅;另一方面,AI技术把研究者从烦琐的基础工作中解放出来,聚焦更深层的学术探索,从而为中华优秀传统文化的传承和创新开辟了更多可能性。这正是技术赋能文化的核心价值。

青年报记者 唐骋华