解码古文字，AI助力古代历史研究

从金融到医学，人工智能（AI）正深刻改变着现代生活。如今，它开始进军古代文本研究：从希腊与拉丁典籍到中国甲骨文，人工神经网络正成为解读古文字的钥匙。它不仅能驾驭浩瀚档案，填补字符空缺，还能解码几乎无迹可寻的罕见或灭绝语言，令古代智慧在现代科技之光下重现辉煌。

2023年10月，费德里卡·尼科拉尔迪收到了一封电子邮件，邮件附带的一张图片彻底改变了她的研究。此图显示了从公元79年维苏威火山浩劫中幸存的一卷莎草纸残骸，它于18世纪在赫库兰尼姆古城的一处豪华别墅遗迹中被发现。这些历经沧桑的莎草纸，曾是数百卷古籍之一，却因岁月侵蚀而变得脆弱不堪，多数已无法展开。

尼科拉尔迪是意大利那不勒斯大学的一名莎草纸学者，她曾参与一项利用AI读取难解文字的研究。而今，她见证了一项奇迹：图片上，一片莎草纸带上，希腊字母密布如织，于幽暗中焕发新生。

这一名为“维苏威挑战”的项目只是AI重塑古代历史研究的“冰山一角”。

神经网络重建古代文本

几十年来，计算机一直被用于对数字化文本进行分类和分析，但目前最令人兴奋的是神经网络的使用。神经网络由相互连接的节点组成的分层结构组成，尤其是具有多个内部层的“深层”神经网络。

卷积神经网络（CNN）模型能够从这些图像中精准捕捉网格状数据结构。CNN模型在光学字符识别领域大放异彩的同时，也开辟了其他多元化的应用途径。例如，中国研究团队在探索甲骨文时，巧妙地运用这些模型来复原遭受严重侵蚀的文字图案，深入分析甲骨文随时间的演变轨迹，并将破碎的文物碎片重新拼凑起来，重现历史原貌。

与此同时，循环神经网络（RNN）作为一种专为处理线性序列数据设计的模型，开始展现出在搜索、翻译以及填补已转录古代文本缺失内容方面的巨大潜力。RNN已被用于为古巴比伦时期数百份格式严谨的行政和法律文本提供缺失字符的智能化建议。

那么，神经网络能否在历史的残片中找出人类专家难以发现的联系？2017年，英国牛津大学的一项合作开启了探索之旅，当时，两名研究人员正面临破解西西里希腊铭文的难题。

古典学者通常依赖对现存文本的理解来诠释新材料，但难以全面掌握所有相关资料。牛津大学研究人员认为，这正是机器学习可发挥作用的领域。他们使用基于RNN的Pythia模型，并用数万份希腊铭文来训练它，最终成功预测了文本中缺失的单词和字符。

2022年，他们又推出Ithaca模型，不仅能预测缺失内容，还能为未知文本提供日期和来源地建议。Ithaca利用了Transformer模型的突破，能捕捉更复杂的语言模式。当前风靡全球的聊天机器人，如OpenAI的ChatGPT就是基于Transformer模型。

翻译复原浩瀚历史档案

韩国研究人员有一项棘手的任务：整理世界上规模最大的历史档案之一。该档案详细记录了27位朝鲜王国国王自14世纪至20世纪初统治时期的日常，涵盖数十万篇文章。美国纽约大学机器翻译专家金亨俊表示，这些文本数据量极为庞大。

将这些文本人工译成现代韩文，预计需耗时数十年。金亨俊携手韩国同行，利用Transformer网络训练自动翻译系统。结果显示，AI译文在准确性和可读性上远超古韩文，有时甚至优于现代韩文。

对于仅存少量文本的古代语言，研究人员也会采用神经网络进行破解。希腊帕特拉斯大学的卡特里娜·帕帕瓦西里欧及其团队，利用RNN恢复了克里特岛诺索斯迈锡尼泥板中缺失的线性文字B文本。测试显示，模型预测准确性高，且常与人类专家建议相符。

面临验证与利用双重挑战

利用AI破解古文字依旧面临诸多挑战。AI技术使非专业人士也能接触到大量古代文献，如何确保研究成果准确无误，成为了首要挑战。神经网络的强大虽令人瞩目，但其偶尔产生的误导性结果，即“幻觉现象”，也让人对结果的可靠性产生担忧。

英国《自然》杂志指出，为解决这一问题，人文科学专家与计算机科学家需携手合作，共同研究并验证AI的解读结果。同时，提倡将所有相关数据（包括原始文本、扫描文件、训练模型及算法）实行开源，以此提升研究的透明度与可验证性。这一做法被称为“数字来源链”，旨在构建一个从原始数据到最终结论的完整链条，便于任何人回溯并核实研究过程。

此外，随着数字化文本数量的激增，如何有效利用这些庞大的数据资源，从中提炼出关于古代社会的重要信息，也是研究人员面临的新课题。这要求研究者转变视角，从单一的文本分析转向对整体文化的深入理解，并尝试将不同地域、不同时期的文本数据相互关联，以获得更为全面的认识。

来源：新华网

“碎片集”项目正在将数以万计的楔形文字数字化。图为：一份天文学文本。 图片来源：英国《自然》杂志

AI生成的“垃圾科学”正侵蚀谷歌学术平台

谷歌学术平台中发现了上百篇疑似由AI生成的文章。 图片来源：瑞典布罗斯大学学院

瑞典布罗斯大学学院研究人员在一项最新研究中警告：无论是从社会知识层面，还是从公众对科学的信任度来看，人工智能（AI）生成的研究都已构成一种威胁。他们最近在文献索引数据库——谷歌学术平台中发现了上百篇疑似由AI生成的文章。相关研究报告日前发表在哈佛肯尼迪学院《错误信息评论》期刊上。

研究人员认为，由AI生成的“垃圾科学”侵蚀学术平台，意味着恶意行为者能以更低成本制造和传播虚假科学，对社会和科研界来说都是一种危险情况。

该研究发现，AI生成的研究带来的主要担忧之一，是证据篡改风险增加，即虚假研究可能被用于战略操纵。此次发现，这些有问题的文章已经传播到网络研究基础设施的多个组成部分，渗透进了各种档案库、社交媒体平台以及其他相关网络渠道。由于传播速度很快，且在谷歌学术平台处于公开状态，即使文章被撤回，也有可能已经传播开来，造成潜在影响。

此外，AI生成的研究也给已经压力重重的同行评审系统带来问题。AI生成的研究在搜索引擎数据库中传播，对参与在线研究的人员的信息素养提出了更高要求，否则，人们很有可能基于错误的信息作出决策。这既是科学不端行为的问题，也是媒体和信息素养的问题。

研究人员强调，谷歌学术平台并不等同于专业的学术数据库。尽管它使用便捷、搜索迅速，但缺少必要的质量保证流程。这一缺陷在普通的谷歌搜索结果中已显现出问题，而当涉及科学知识的普及与传播时，这一问题则变得更为严峻和复杂。

来源：科技日报