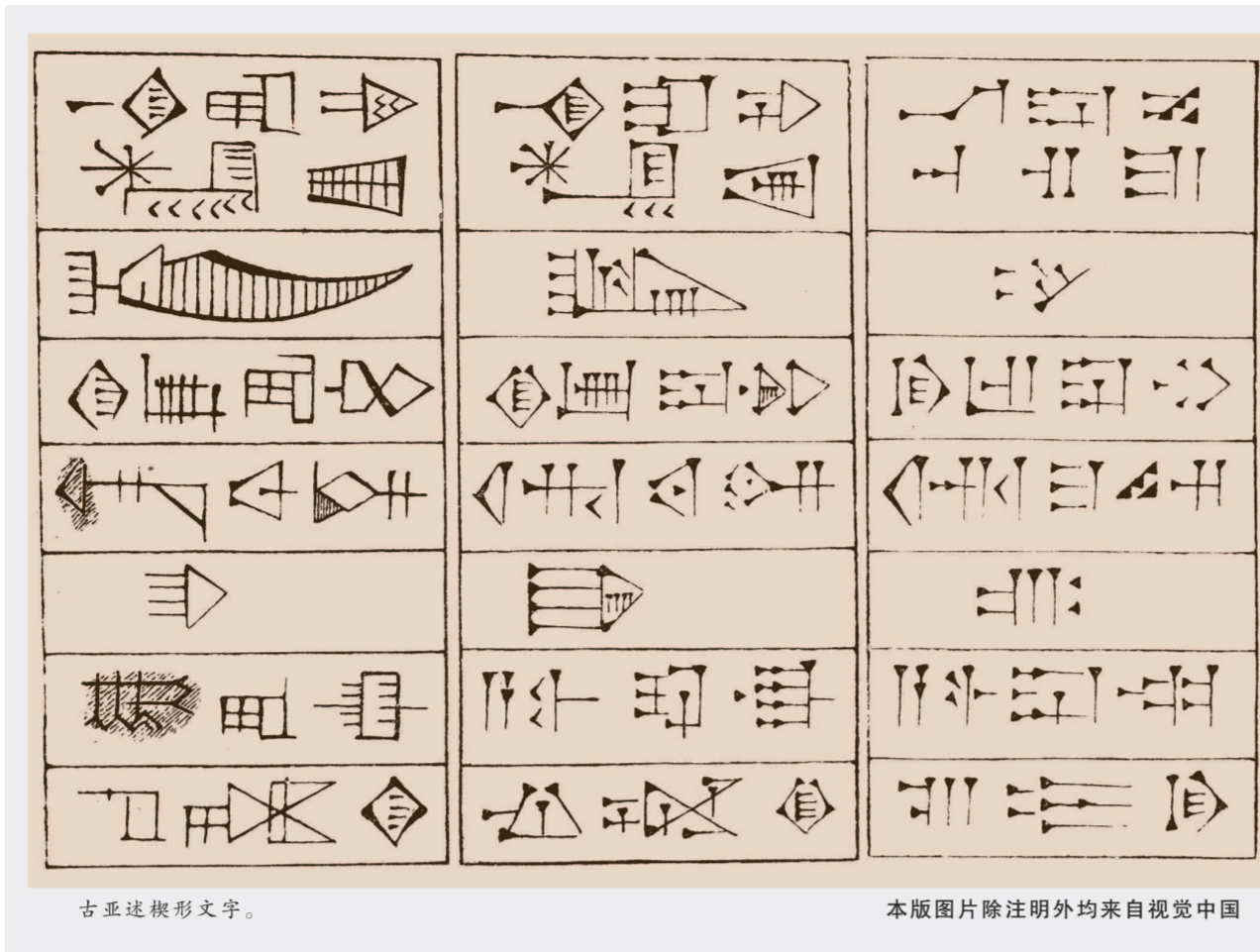


从科学发现到艺术创作,创造性人工智能迎来飞速发展

AI破译楔形文字,探索“世界第一文明”

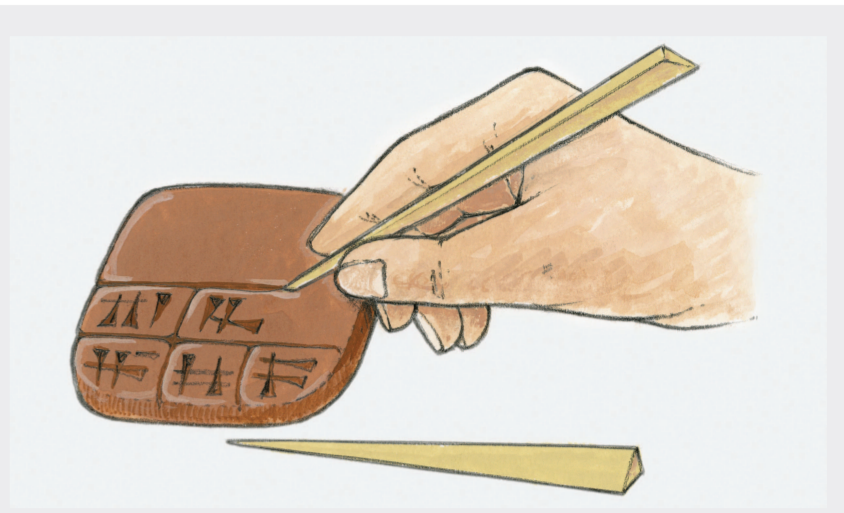
无论是科学发现还是艺术创作,人工智能(AI)正向着原来被认为只属于人类的领域进军。在《科学》杂志日前评选出的2022年度十大科学突破中,“创造性人工智能的快速发展”备受关注。从预测蛋白质结构到设计新药、疫苗,AI都取得了“攻城略地”式的进步,甚至在人文艺术领域频频展现高光。

楔形文字是人类最古老的文字之一,揭开它的神秘面纱,一直是研究人员孜孜以求的目标。如今,人工智能正在帮助破译美索不达米亚人刻写在陶泥板上的楔形文字,这些古老的泥板文书将为我们揭开人类历史早期一些不可思议的故事。



古亚述楔形文字。

本版图片除注明外均来自视觉中国



用芦苇笔在泥板上书写苏美尔楔形文字。



保存在大英博物馆中的楔形文字泥碑。

图/New Scientist



人工智能对楔形文字泥碑进行扫描翻译示意图。

图/New Scientist

楔形文字的起源与演变

楔形文字的故事可追溯到6000年前。美索不达米亚地区,那里是位于底格里斯河和幼发拉底河之间的一片肥沃地区,位于如今的伊拉克境内。这一时期的人类社会正在经历重大变革,从小规模的农业聚居地逐渐发展成庞大的城市中心。苏美尔人在这建立了最早一批城市,乌鲁克城是其中最具有代表性的重要城邦之一,城中寺庙林立,沟系纵横,到公元前3000年时,城中人口已达5万。城市成为周边地区的行政管理中枢,管理复杂劳务系统的官僚体制也开始形成。

苏美尔语完全不同于如今我们所知晓的任何语言,并且早已消逝在历史的长河中。但苏美尔人留下的大量楔形文字泥板文书,让我们能有机会对当时美苏尔人的社会生活一探究竟。美索不达米亚地区是楔形文字的发源地,当时的人们将芦苇的末端压入尚未变硬的湿软陶泥板上,写下一个个被称为楔形文字的古老符号。

虽然人们通常将楔形文字形成的文化遗产与诗歌和文学联想在一起,但最早的楔形文字的使用完全是出于行政管理的需要,比如作为奴隶转让交割手续的凭证,或接收牲畜的签收收据等。最典型的一个例子是保存在大英博物馆内的一块刻写了啤酒配额记录的泥板文书,画一个瓶子代表一瓶啤酒,在瓶口上画圈代表数量,这是最初的象形符号文字,后来逐渐演变为更为抽象的文字形式。

楔形文字从简单的记事工具演变为一种语言表达形式,经历了相当漫长的时期。最早的王室铭文出现于公元前2700年左右,最早的文学文本出现于更晚的一百多年后。最早的文学作品作者是生活在4300年前的一位公主、女

祭司和诗人恩赫杜安娜。她写下了许多赞美诗和神话故事,其中最著名的是吉尔伽美什史诗,关于一位国王追求永生的故事。

美索不达米亚古代文明的影响一直延伸并渗透到现代人类的生活中,圣经中的一些故事与苏美尔人的文化传统甚有渊源,苏美尔人还是如今仍用于钟表中的六十进制计数系统的发明者。

破译楔形文字困难重重

楔形文字本身并不是一种语言,而是一种书写符号系统,好比英语中用的字母,同样也用于法文和德文中一样。苏美尔人从世界上消失之后,他们的楔形文字书写方式流传了下来,成为许多其他语言的书写符号,如阿卡德语、赫梯语、古埃及西亚语等。在使用了3000年之后,这些书写符号与语言,以及描述了历代王国兴衰的记录,一起消失在了历史的长河中。

我们如今能够了解这段遥远的历史,要归功于用来书写楔形文字的陶泥板。这是当时的一种廉价材料,随处可见,并且可以耐久保存。“非常幸运的是,除非被扔进河里,或被完全碾碎,这种上面写有文字的泥碑都能得以保存下来。” 苏美尔语说。

世界各地发现的泥碑数量已有成千上万,是我们探索早期人类文明的珍贵文化遗产,其中包括地球上最早帝国的编年史,苏美尔人创作的赞美诗、来往信件、购物清单,甚至还有客户的投诉信。“有人说,人类历史的前半部都记录在了这些楔形文字泥板上。” 德国慕尼黑大学的恩里克·杰梅内兹说道。

这些泥碑中隐藏着的远古时代的秘密不断被破译。2017年,学者发现,一块被称为“普林顿322”的3700年前的泥板文书是世界上最早出现的三角学知识。它曾被一名叫普林顿的人收

而得名,“322”是普林顿的收藏编号,但其最初来源不详。碑文显示,生活在美索不达米亚中部和南部讲阿卡德语的巴比伦人,是最早研究三角学的人。

2021年,对1894年在伊拉克出土的一块泥碑的新的分析研究发现,早在毕达哥拉斯(古希腊哲学家、数学家)出生几个世纪之前,巴比伦人就已开始对三角学进行计算和研究。遗憾的是,如今世界上只有75个人能够流利阅读楔形文字,大多数泥板文书只能束之高阁,在博物馆里静静蒙尘,鲜为人知。楔形文字太复杂,太难读懂,长期以来一直困扰着人们。“碑文中文字的含义往往模棱两可,而一个词语可能会有不止一种书写方式。”杰梅内兹说。还有一个难题是,很多楔形文字泥碑并不完整,不是破裂缺角,就是碎成多个小块,通常边缘部分还有磨损,导致一段记事不是没有开头,就是没有结尾,或是中间部分缺失。

亚述帝国国王阿什巴尼帕尔的藏书室是世界上现存最古老的皇家图书馆,收藏了来自美索不达米亚各地的大量书写文献,3万块泥碑记载了各种文献资料,包括宗教仪式、医学百科全书、天文观测,以及王室开拓疆域的功绩等。可惜的是,公元前612年,在城市被攻陷洗劫中,这座古老的皇家图书馆被焚为废墟。大英博物馆保存的大火中幸存下来的一些残碑上面,焦黑的灼烧痕迹仍然历历在目。

要想还原历史,就要将这些碎片像做拼图游戏一样拼合起来,没有任何参考依据,难度很大。而且记载有楔形文字的泥碑分散在世界各处,收集起来也很困难。例如,吉尔伽美什史诗的开头部分保存在一块很小的陶泥碑文书残片上,光是找到并确认这块泥碑,就花了一百多年时间。如今,有了人工智能来帮助做这项工作,一切都将有所不同。

AI训练“复原”巴比伦文学

“残片数据库”是杰梅内兹于2018年设立的“电子巴比伦文学项目”的一部分。该项目利用人工智能找出哪些残片同属于某一块泥碑,然后将它们重新组合起来,以还原阿什巴尼帕尔的皇家图书馆残存的泥板文书和其他泥碑的内容。杰梅内兹正在用开发出来的算法对楔形文字的不同变体,以及多种副本之间的微小差别进行比较研究。

经过训练的AI可以对碑文进行翻译。研究人员根据文本符号的音译,对人工智能进行深度学习训练。楔形文字符号的发音依据是拉丁字母的发音方式,就像汉字的普通话发音用的是拼音一样。经过学习训练的人工智能可以预测哪些楔形符号可能出现在缺失的残片中,还可以在数字化的庞大残片数据库中搜索特定的楔形文字符号。2019年,人工智能帮助在多个残片中找到了吉尔伽美什史诗中的缺失部分,还发现了一种新的古代文字体,其特点是滑稽、夸张和模仿,比如一堆会开玩笑的驴粪,主要是用来帮助小学生学习读写技能的。

杰梅内兹与伊拉克巴格达大学伊尔克分校的安玛·法迪尔一起,通过人工智能的帮助,发现了另一种以前不为人知的文字体——城市赞美诗。比如,巴比伦城赞美诗对寺庙僧侣的日常进行了细致的描绘。2021年,在世界上首次使用人工智能完全自主识别楔形文字片段的过程中,研究人员发现了名诗《正义的受难者》中缺失的片段。

其他一些研究人员关注的重点是反映普通行政工作的楔形文字文本。研究发现,大量内容短小的收据文本出自不同的机构,如寺庙或当地统治者的官

邸,或是个人(如商人)之间进行交易的书面文本。这些丰富的信息是美索不达米亚古老文明的重要组成部分之一。

苏美尔泥碑文本中通常包含个人姓名和日期,根据这些线索,可以追溯了解某人在当时社会中承担的角色。例如,形成于公元前2300年左右,被称为“Mama-ummi档案”的80块泥碑文本描述了一位名叫Mama ummi的女主管,她负责一个由180名织工组成的团队。这份资料表明,当时的女性有很多的工作机会,这一点非常出人意料。

面对浩如烟海的楔形文字行政信息,加拿大多伦多大学的希瑟·贝克利于2017年建立了完美的楔形文字机器翻译和自动分析系统。在不久前的测试中,他们先用45500个音译短语对不同的人工智能算法进行训练,每个短语由多达19个单词组成,然后对人工智能将苏美尔语单词翻译成英语的能力进行了测试。2021年公布的结果表明,某种特定人工智能算法的翻译准确性可达95%。该系统还能从文本中提取关键词信息,对人员、地点和神祇等类别进行识别。

2021年,以色列耶路撒冷希伯来大学计算机科学家加布里埃尔·斯坦诺夫斯基和他的同事还找到了一种预测残片缺失文字的方法,类似于在手机上输入文字时的自动预测。

研究人员尝试给深度学习人工智能喂入上万块阿卡德语书写的楔形文字泥碑的音译单词信息,发现人工智能根据上下文填补缺失部分内容的准确率能达到80%。

人工智能的另一项潜在应用是确定泥板文书的年代,即以已知年代为线索,训练人工智能算法预测其他文件中缺失年代信息的能力。

但是,从符号音译来解读楔形文字是一回事,要解读楔形文字本身又是另一回事。楔形文字不仅随着时间的推移在演变,其拼写也在变化,不同时期的楔形文字还被用于不同的语言。最棘手的是,符号之间没有间隙,要推断出哪组楔形符号构成某个单词非常困难。拉丁字母包含26个字母,但楔形文字符号多达900多个,而且它们看起来非常相似。

为克服重重困难,研究人员正在开发类似于文本识别的人工智能计算机系统,向破译楔形文字的目标迈进。研究人员训练了一种叫做“DeepScribe”的机器学习系统,用来解读波斯波利斯卫城档案的数千块楔形文字文本,这是在大约公元前500年前修筑的一道防御工事墙中发现的埃兰语行政管理文书。研究团队成员、美国芝加哥大学的苏珊

娜·保卢斯指出,“计算机能够识别单词的起始符号和终结符号的边界,然后用方框围住,这是一项巨大的成就。对于那些没有分隔的一连串符号,人们通常没有办法知道一个单词的起始位置和终结位置。”

未来有望实现AI即时翻译

在人工智能的帮助下,未来将有望实现将符号识别系统和现代语言翻译系统一体化的目标。这将意味着也许有一天,我们就可以用手机拍下博物馆里一块泥碑文书的照片,并即时读出上面的内容。要实现这一宏大目标,需要一个庞大的数字化文本数据库来支持人工智能算法的训练和学习。目前世界各大博物馆中约50万件楔形文字文本中,只有一半经过了音译翻译或文本翻译,经数字化处理的文本只有大约10万件。“楔形文字数字图书馆计划”和“电子巴比伦文学项目”的进展正在改变目前这一现状。

以色列阿里尔大学的萨伊·戈尔丁表示,“楔形文字文本数字化可为学者带来大量新信息和新联系。一旦能够将这些信息放在一个庞大的神经网络中,下一个重大突破的契机就会到来。我们就可以对那个时代古人的生活有一个全面的了解。”

开始阶段需要大量辛苦的工作,首先要获得各地博物馆和私人藏品中所有楔形文字泥碑文书的高清图像。目前,作为“电子巴比伦文学项目”的组成部分之一,大英博物馆收藏的阿什巴尼帕尔图书馆4万件楔形文字泥碑残片的拍摄工作正在进行中。

在一间特殊的拍摄工作室里,摄影师阿尔伯特·吉亚尼斯给每件楔形文字泥碑拍摄六张图像——正面、背面、顶部、底部和两个侧面。吉亚尼斯说,文本并不总在边缘处结束,有时甚至区分前后和上下都很难。

然后,计算机软件会将这六张图像的内容拼接组合起来,再由杰梅内兹这样的楔形文字专家进行辨认和翻译。据预计,2023年将完成所有楔形文字藏品的拍摄工作,并向大众公开。以往深藏于大英博物馆内的楔形文字藏品不是普通大众都有机会观赏到的,但经过数字化处理的藏品,可以让人们舒服地坐在自己家中,欣赏苏美尔阿什巴尼帕尔图书馆的泥碑文献和其他早已失传的楔形文字文本。

回望历史,数千年前的人类给我们留下了卷帙浩繁的楔形文字文书,以及承载其中的文化思想;展望未来,这些独具远古文化特色的泥碑文本,在人工智能的强大计算能力下,重新组合和解码,以现代科技的数字化形式,重现古代人类文明曾经的辉煌。



相关链接 解读楔形文字符号

如果不是在伊朗人迹罕至的高山山顶上发现的贝希斯敦铭文(上图),楔形文字符号也许至今未能破译。这块刻有三种文字的陶泥碑是楔形文字版的“罗塞塔石碑”。罗塞塔石碑同样刻有三种文字,是破解古埃及象形文字的一把钥匙。

贝希斯敦铭文刻于公元前520

年左右,为纪念波斯王大流士平乱胜利,用楔形文字符号所刻,用了三种不同的语言。1764年,西方探险家在伊朗的山顶上踩着摇摇晃晃的梯子,首次拓下了这篇铭文的拓本,但当时无人能够破译这些文字。

碑文用的三种语言文字中,最先破解的是古波斯语。从大流士等名字

的书写方式来看,碑文上的另一种语言阿卡德语也是可以破译的。阿卡德语是在古代美索不达米亚使用过但今已灭绝的一种语言。“阿卡德语的秘密正渐渐被解开,如果没有那把‘钥匙’,我想也许永远无法揭开这个秘密。”伦敦大英博物馆的欧文·芬克尔说道。

这是个很好的突破口,以此为契机,破解最原始的楔形文字苏美尔语,相对就容易了。楔形文字的铭文多为双语文本,阿卡德语刻写的铭文下面通常会加上苏美尔语的译文。