机器学习如何破译早已消亡的古老语言？

人工智能4 月 6 日

在大英博物馆(British Museum)的柔光照射下，人们只能勉强看到镌刻在这些古老泥板上的密密麻麻的楔形标记。这些细小的标记是世界上最古老的书写系统——楔形文字的遗迹。

楔形文字起源于 5000 多年前的美索不达米亚，位于底格里斯河和幼发拉底河之间，也就是现在的伊拉克。楔形文字捕捉了一个长达 3000 年之久的、复杂而迷人的文明。从王室兄弟姐妹之间愤怒内斗的信件，到安抚一个任性婴儿的仪式，这些石碑让人们可以从另一个独特的视角了解历史初期的社会。

它们记录了阿卡德、亚述和巴比伦帝国的兴衰，这是世界上第一个帝国。据估计，人们已经挖掘出了约 50 万块楔形文字板，但还有很多仍深埋地下。

约 150 年前，学者首次破译楔形文字。然而，只有一小部分能读懂这种文字的人才了解其中的秘密。目前，仍有约 90%的楔形文字未被翻译出来。

但是，这种情况可能会有所改变，这都要归功于现代工具——机器翻译。

"人们并不了解美索不达米亚文明对自身文化的影响，"多伦多大学亚述学研究员佩龙(Emilie Page-Perron)说。美索不达米亚文明孕育了车轮、天文学、一小时 60 分钟的计时制、地图、洪水和方舟的故事、以及第一部文学作品——《吉尔伽美什史诗》。这本诗集主要是用苏美尔语和阿卡德语写成的，能读懂这些语言的学者少之又少。

佩龙现在正在进行的一个项目，是用机器翻译公元前 21 世纪以来美索不达米亚文明的行政记录，数量多达 69000 份，其目的之一是为新的研究发掘过去。

佩龙说："我们虽然已经获得了关于美索不达米亚人生活的信息，但却没有真正从（美索不达米亚）不同领域专业人士的知识中获益，比如经济和政治领域。如果有渠道（了解这些知识），我们能更好地了解那些古老的社会。"

除了石碑，还有 5 万多枚美索不达米亚雕刻印章散落在世界各地。几千年来，美索不达米亚人使用由雕刻石头制成的印章，这些印章被压入潮湿的粘土中，用来标记门、罐子、石板和其他物品。这些刻章中只有十分之一被编入目录，更不用说翻译了。

牛津大学亚述学教授达尔(Jacob Dahl)表示:"我们所获得的关于美索不达米亚文明的资料比希腊、罗马和古埃及的加起来还要多,但真正的挑战在于找到能读懂它们的人。"

佩龙和她的团队正在对一个数字化数据库中的 4000 个古代行政文本样本编写算法。这些行政文本包括交易和运输记录,比如把羊、芦苇束或啤酒运到寺庙或个人手中的记录。这些文字最初是用芦苇笔刻在粘土上的,现在,学者已经把它们音译成了我们的字母表。例如,苏美尔语中表示"大"的词可以写成楔形文字,也可以写成英文字母表中的"gal"。

这些行政文书的措辞很简单。例如,"第 15 天,厨房有 11 只母山羊"。这种特点使得它们特别适合被自动化处理。一旦算法学会了将样本文本翻译成英语,它们就能自动翻译其他经过音译的石碑。

佩龙表示:"如果单独看我们正在研究的文本,它并没有那么有趣。但如果你把它们当作一组文本来看,就有意思多了。"她预计英文版平台将在明年内上线。这些记录向我们展示了古代美索不达米亚人的日常生活,包括权力结构和贸易网络,同时还展示了社会历史的其他方面,如女工的角色。平台上可被检索的翻译,将使不同地方的研究人员都能探索到古代生活的丰富面向。

佩龙解释说:"这些人与我们是如此不同,但他们也面对着和我们一样的基本问题。理解美索不达米亚文明,能够帮助我们理解生而为人的意义。"

她希望机器分析也能弄清苏美尔人的一些特征，这是至今仍困扰着现代学术界的难题。这种已经灭绝的语言与任何现代语言都没有联系，但却保存在以楔形文字书写的碑文中。这可能是我们与更古老，甚至没有历史记载的社会之间最后的联系。

"苏美尔语可能是数千年前的语言大家庭中的最后一个成员，"芬克尔(Irving Finkel)说。"文字及时地出现在这个世界上，拯救了苏美尔语……幸运的是，在苏美尔语与其他文字一起消失之前，我们及时地开始学习这种语言。"

Image caption 能够识别古代石碑文字的算法能够帮助研究人员将它们与制造它们的原始石印进行匹配。

芬克尔是世界上顶尖的楔形文字专家之一。他在大英博物馆堆满书的办公室里讲解了手稿是如何慢慢被破译的，这多亏了一位国王的多语种铭文，就像罗塞塔石碑帮助研究人员理解了埃及象形文字一样。

他说:"当你与千年前的灵魂进行交谈时你会惊讶地发现，这简直太有趣了，仿佛在和他们打电话。认识他们是世界上最令人兴奋的事情。"

触碰古老宝藏

只有少数人能接触到拥有 5000 年历史的石碑，但多亏了先进的成像技术，现在任何人只要能上网就能接触到这些宝藏。比如，世界上现存最古老的皇家图书馆，人们正在将它数字化。这座图书馆位于尼尼微，由亚述国王亚述巴尼帕（Ashurbanipal）建造。大英博物馆展出了图书馆里幸存的一些碑文，是亚述巴尼帕专题展览的一部分。虽然早在公元前 612 年，尼尼微遭遇洗劫时，这些碑文被火烤得又黑又硬，但上面得文字仍可辨认。

新的成像技术让人们在处理这些古老且破损严重的文本时更加轻松。有了精细的图像，人们就有可能找出那些肉眼看不见的模糊标记。

达尔和他的同事一直在进行一个名为"楔形文字数字图书馆倡议"(Cuneiform Digital Library Initiative)的项目，将储存在德黑兰、巴黎和牛津馆藏中的碑文及印章进行数字化处理。这个庞大的在线数据库已经包含了世界上约三分之一的楔形文字，以及一

些未被破译的书面语言，如古伊朗的原始埃兰语。如果没有这样庞大的数字资源，让机器进行翻译几乎是不可能的。

Image caption 人们在先进的成像技术及机器视觉工具的帮助下破译古代语言，如原始埃兰语。

数字化还帮助研究者们将散落在世界各地的文本拼凑起来。达尔与南安普顿大学及巴黎南泰尔大学的研究者一同对美索不达米亚的 200 多枚石印的 3D 图像进行了数字化处理。在试点项目中，他们使用了人工智能算法校验了 6 块碑文，并识别出在世界其他地方发现的与之匹配的石印。算法准确地挑选出了两块现存于意大利和美国的石碑，这两块石碑上盖的石印是一样的。

在过去，想要将石印和印痕匹配起来困难重重，因为许多石印储存在数千英里之外的地方。达尔预计，五年内可以将所有的印章进行数字化处理，这样就可以追踪其他方面的信息。比如说，有迹象表明，某种石头更受到女性的青睐。

达尔说："要得出这种结论必须拥有大量经过处理的石印图像，并运用算法和机器学习等技术。"他希望，人工智能的发展能帮助探索世界各地收藏品中蕴藏的丰富信息。

"亚述研究涵盖了人类历史的一半，是一种濒临灭绝的文化遗产。我希望亚述学能走在这方面的前沿。"

破译古人的语言

成像技术也改变了对于未破译文本的研究。对于数量少、具创造性文本的破译，人类往往比机器做得更好，人类有着对生活和组织方式的深入理解，以及高度的灵活性。

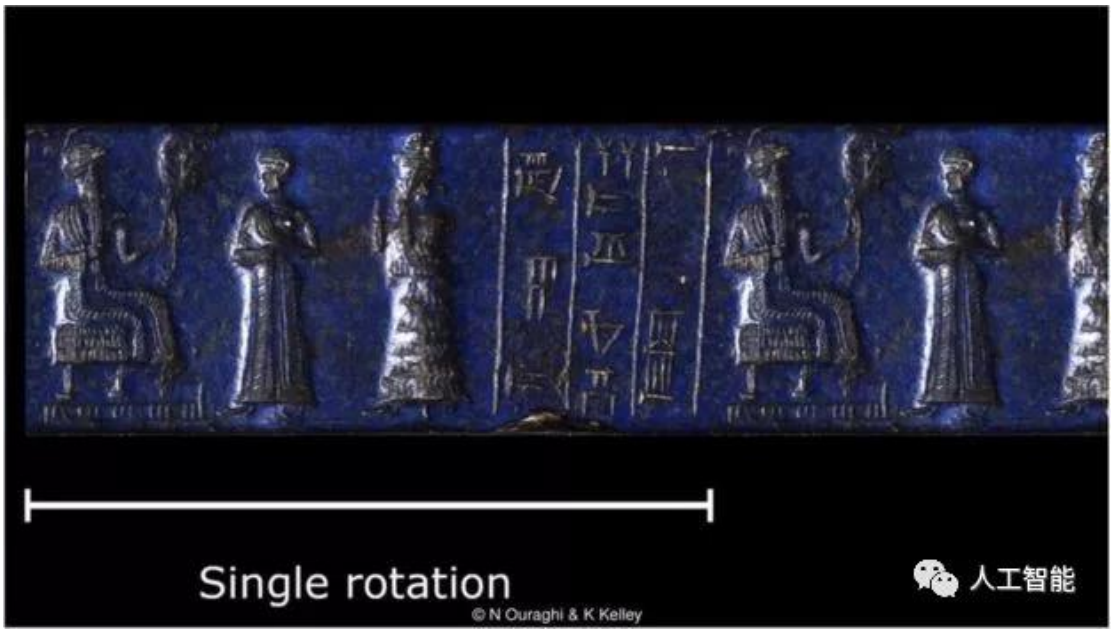

Image caption 三维成像技术能够详细检测青金石石印这样的圆柱形石印。

例如，早期的楔形文字符号并不是线性排布的，而是简单地与画在周围的方框排在一起。原始埃兰语是三维立体的，一个圆印的深浅不同意义也不同。但是，技术可以放大、分享和比较图片的细节，加快了破译进程。

一直致力于破译神秘文本的达尔说："获得正确的图像是问题的核心。原始埃兰语研究缺乏的正是这个。"

这些进步已经超越了亚述学领域。剑桥大学高级研究员斯蒂尔（Philippa Steele）是研究古克里特和希腊早期文字系统的专家。其中包括"线形文字 A"(一种未破译的文字)和"线形文字 B"(一种古代希腊语的书写形式)。

归功于成熟的成像技术，古代石碑上的文字被很好第呈现，斯蒂尔才在其中发现了新的细节。

她说:"你可以辨认出肉眼很难辨认的特征。"这些特征通常与撰写文本的人与文本交互的方式相对应。例如，对于线性 B，你可以分辨出更改的痕迹。有时你可以判断出撰写这份文件的人是什么时候想出来了什么，然后又在上面写了什么。

Image caption 伊拉克考古学家发掘出数千块刻有世界上最古老文字的石碑。

佩龙希望机器最终能够翻译更复杂的苏美尔语石碑和其他语言 ,比如阿卡德语。她说:" 关于古代文化，还有很多东西有待发现。 "

也许有一天，我们将能够阅读所有古老文字的翻译版本，尽管当我们去世时，美索不达米亚的许多未解之谜还未解开，尤其是现在许多缺失的楔形文字碎片仍深埋地下，等待挖掘。

古代美索不达米亚的国王们深深地思考着过去和未来。他们崇敬前朝的楔形文字，将记录着他们的名字和成就的铭文埋藏地下，寄望后世的统治者会将荣耀归于自己。

在某种程度上，他们的愿望已经实现。他们的经历过的战争和征服可能已经被大多数人遗忘，但是他们最强大的发明——文字——在过去的几千年里助力了人类思想和技术的发展。而现在，人类开始训练机器从过去中学习。

- END -