学界公认甲骨文有4500多个单字，迄今已破译近1500字，剩下3000多字都较难释读

破译甲骨文，AI准备好了吗

本报记者 **沈竹士**

7月5日，上海，2024世界人工智能大会。安阳师范学院团队宣布全球首个甲骨文多模态数据集正式开源。所谓多模态，是指包含一万片甲骨拓片、摹本，以及甲骨文单字对应位置、对应字头、对应隶定字以及辞例分组、释读顺序等数据。研究人员可基于该数据集开发甲骨文检测、识别、摹本生成、字形匹配以及释读等方向的智能算法。

近年来，安阳师范学院甲骨文信息处理教育部重点实验室利用计算机缀合甲骨碎片图像70余组，位列全国第一。其中一组缀合后形成了新的连贯文辞，如果释读无误的话，这句话可能记录了公元前1900多年的一次日偏食天象。这引起人们的极大关注。

最近二十年，甲骨文破译逐渐进入瓶颈期。为此，在政府相关部门推动下，多所高校研究团队致力于探索人工智能（AI）辅助研究甲骨文的技术。国内互联网巨头和科技公司纷纷入局，与学术界开展跨学科合作。人工智能的应用为甲骨文研究提供了新的思路。玄幻的殷商甲骨文与科幻的人工智能碰撞，这是属于中华文化独有的浪漫。

俄罗斯国立爱米塔什博物馆所藏甲骨缀合结果。原文为“壬辰王卜，贞王其若…呼比𠂤眾…其二人異史…”。释文为“壬辰日王占卜，贞问，王赦免亘方战俘，令其配合比𠂤眾…二人去做某事。”

北珍435　　北珍438

两片来自北京大学珍藏甲骨文字的碎片经计算机辅助缀合的结果。缀合前卜辞分别为1.丙戌日又□2.即□王卜曰，叀王求，又𢦏。五月。通过缀合发现“日又”与“即”连读。完整卜辞为“丙戌日又即，王卜曰：叀王求，又𢦏。五月”。有学者认为“又”通“有”，“即”通“食”，意为丙戌这天出现日偏食（日有食），商王占卜认为会带来灾咎，于是进行“𢦏”的祭祀。

AI需要一个怎样的甲骨文数据库

投喂给人工智能的标准化、多模态数据集，起点是二十年前一位数学老师开发的输入法。

1991年，安阳殷墟花园庄东地H3坑内出土甲骨1583片，这是殷墟甲骨发掘史上第三次重大发现。彼时，从河南师范大学数学系毕业的刘永革分配到安阳师范专科学校（安阳师范学院前身）任教才第三年。在职业生涯的前十年里，他与甲骨文研究并无交集。

上世纪90年代末，安阳师专安排青年教师进修考研。刘永革等十人来到西安，目标是考上西北工业大学计算机工程学院。当时个人微型计算机刚刚兴起，进机房之前需要穿鞋套以防静电。刘永革是数学专业出身，考试有四门课，其中三门以前没学过，他便去书店买来专业书籍现学。有同学新买了一台照相机，招呼大家去秦始皇陵兵马俑坑参观游览，开玩笑说：“刘永革，别复习了，你陪我去，你肯定考不上嘛。”刘永革应该没去看兵马俑——他在2000年获得计算机软件与理论硕士学位，方向是数据库应用。

21世纪初，安阳师院有一批从事甲骨文研究的中青年学者，包括李雪山、韩江苏等，他们都曾在上世纪80年代“殷商文化研究班”受业于甲骨文专家胡厚宣。写论文要引用甲骨文，甲骨文怎么输入电脑？虽然有一种甲骨文编码输入法，但是学习成本很高，就像五笔字型输入法一样，需要背诵一整套编码。老师们找到已在计算机科学系任教的刘永革，希望他开发一种完全不同以往的新的输入法。首先，用软件描摹甲骨文字，将描出的图形矢量化，制成字体库。再根据日本学者岛邦男的甲骨文部首自然分类法设计检索体系。用Visual C++编写动态数据交换程序。使用时，呈现甲骨文常用部首的图形界面，只需鼠标点选检索，再点选需要的文字即可，不用背码。对于文字数量不多的甲骨文来说，这种输入法是非常合适的。

涉及甲骨文研究，不仅要输入单字，还要能输入整句、要找出前人的释读成果进行对照，最好配上甲骨拓片或摹本的原图。圈内学者常开玩笑说，其他学科阅读资料可以用文本文档或者word文档，甲骨文研究只能看PDF文件——用它才能浏览清晰的拓片图像。历史与文博学院的韩江苏教授意识到，甲骨文研究需要一个字、图、文资料一体化、便于检索的数据库。“甲骨文图文资料库”2004年成功申请国家社科基金，甲骨文输入法的成功经验在焉，刘永革很自然地加入了课题组。好好一个计算机专业老师，毅然跨界投身甲骨文的世界。他带领计算机系的年轻人从头学习甲骨文，为课题组增添新鲜血液。至结项验收时，收录数十种权威研究文献的精华和7万多张甲骨拓片。

郭青萍是安阳师院中文系教授，退休后自学甲骨文并从事甲骨文篆刻。一次，他请刘永革帮忙检索几个现代汉字对应的甲骨文字形。刘永革很快把结果给到了老先生。“我翻书查找可能要花一个月，你这么快就找到了?！这个电脑很好。我也要学电脑!”那年郭青萍89岁，家里人不支持他。他拿出7000元偷偷交给刘永革，要他帮忙选购一台电脑。刘永革说：“老先生好学呀。我给他买了一台显示器很大的那种，方便他看字。后来他用电脑又写了三本甲骨文方面的书稿交付出版。”2008年，刘永革等申报的《基于甲骨文语料库的计算机辅助考释技术研究》获批国家自然科学基金项目。甲骨文资料的数字化极大地便利了研究者，也为即将到来的人工智能时代做好了铺垫。

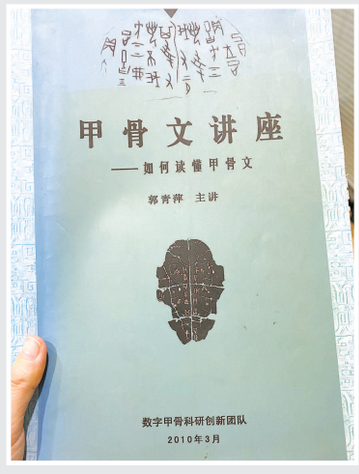
2016年3月，谷歌旗下DeepMind团队开发的AlphaGo（初级围棋）程序击败韩国九段棋手李世石，震惊世界。这也被认为是一个人工智能发展大周期的元年。一个月后，国家相关部委领导在河南安阳调研时说，要利用大数据、云计算等现代技术手段做好甲骨文的破译工作。两年后，安阳师院甲骨文信息处理教育部重点实验室获批，刘永革出任实验室主任。按照规定，教育部重点实验室学术委员会主任应由院士担任。“我们安阳是小地方，哪认识什么院士哟。”

好在，他们“蹲”到了2015年新当选中国工程院院士的戴琼海。刘永革就聘请他做学术委员会主任。戴琼海是清华大学自动化系教授，长期致力于立体视觉和计算摄像理论、关键技术研究，现任中国人工智能学会理事长。

2019年是甲骨文发现120周年。安阳师院在甲骨文研究专家宋镇豪指导下，发布“殷契文渊”甲骨文数据平台。这是当今世界资料最齐全、最规范、最权威的甲骨文数据平台，对国内外研究者免费开放，至今已更新4期，包括甲骨著录154种、甲骨论著34417种，收录23余万种图像。利用上亿像素的照相机，通过高清拍摄、微距拍摄、三维建模、红外线拍摄、多光谱拍摄，对每一片甲骨拍摄150余张照片。借助微痕增强技术，使研究者能清晰地看到甲骨上较浅的刻痕，更准确地分析笔画和轮廓。部分有特殊含义的甲骨文是用丹砂“涂朱”的，对这部分文字的研究也是甲骨文研究中的一个分支。有些甲骨因年代久远，红色丹砂脱落殆尽，但通过光谱分析，仍然可以确定甲骨文中的涂朱部分。此外，根据机器学习的要求，添加图像数据标注。

回首过去，当初为甲骨文输入法制作的矢量字库，已经“魔改”得面目全非。技术发展超越人的想象。

甲骨文专家胡厚宣上世纪80年代在安阳师专主讲“殷商文化研究班”，培养了一批甲骨文人才。



安阳师范学院计算机系青年教师跟随中文系教授郭青萍学习甲骨文所用的讲义。

张展操作计算机辅助甲骨碎片缀合程序，检查运行结果。　　沈竹士摄

用计算机把破碎的甲骨拼起来

张展，2019年获中国科学院大学计算机应用技术专业博士学位，2021年中国科学院沈阳自动化研究所博士后出站，研究方向为计算机视觉、模式识别和数字图像处理。在学校，张展和河南安阳籍郭安是室友。找工作时，郭安回安阳师院谋求教职，张展便顺道来安阳看看机会。这是他第一次见到刘永革。老刘希望张展留下，但后者还没有打定主意。一段时间后，刘永革接到张展的电话。他非常高兴：“不用说相应的待遇。张展来我这里，他成家前，我发动系里的老师，要求每个人都给他介绍对象。我得让他留在安阳。”

刘永革一眼看中张展，因为他的研究方向非常适合从事计算机辅助甲骨碎片缀合工作。而甲骨碎片缀合是短期内能够实质性推动甲骨文释读破译的手段。

当前的甲骨文释读工作确实处在一个瓶颈期，在甲骨学再继续发展的道路上，遇到了文字释读滞后的障碍，给甲骨学商史研究的再深入造成了困难。中国文字博物馆于2016年至2024年间开展了两次甲骨文释读优秀成果征集评选。对破译未释读甲骨文并经专家委员会鉴定通过的研究成果，单字奖励10万元。第一次，复旦大学蒋玉斌摘得一等奖，拿到了10万元奖励。第二次，复旦大学陈剑和吉林大学周忠兵同时获得一等奖。8年，3个字，这就是今天破译甲骨文的速度。

为何破译如此之难？目前学界公认甲骨文有4500多个单字，其中已经破译近1500字，剩下的3000多字都是较难释读的，譬如没有对应的现代汉字，或是后世不再使用的地名、人名。甲骨文破译是从已知推理未知。1991年安阳殷墟花园庄东地H3坑的发掘是距今最后一次甲骨大发现。近三十年，新发现的甲骨增量太少，也使释读研究工作陷入巧妇难为无米之炊的境地。

为此，很多学者将目光投向存量甲骨挖潜，希望从中压榨出有价值的新线索。现存甲骨多以碎片的形态存世。一是因为甲骨用于占卜，经过钻凿、火烧，沧海桑田，绳编断绝。二是早期甲骨收藏者在安阳小屯村收购甲骨，按片计价。村民便将挖出来的甲骨掰碎出售。所以后来就改为按甲骨上有多少字来计价收购。甲骨往往沿龟腹甲的天然纹裂而碎，其小者比人的指甲盖大不了多少。如果将甲骨碎片缀合起来，就能得到新的连贯的句子，学者加以句读，从而获得全新的解读。

然而，甲骨整理繁难且极费人工。故宫博物院是世界第三大甲骨收藏单位，所藏2万多片殷墟甲骨，此前绝大多数从未整理出版。“故宫博物院藏古文字数字平台”的阶段性成果，也仅仅是公布了《故宫博物院藏殷墟甲骨文》“马衡卷”“谢伯殳卷”中的300余件甲骨藏品高清影像及其拓本。人工缀合甲骨碎片需要记忆大量的甲骨文信息，专业要求高、工作量大。一所高校能有几个研究甲骨文的人才，他们寒窗苦读，皓首穷经，才堪堪够格参与这项工作。古人考释文字如同射覆，意即如猜谜一般，靠直觉，没有数学公式推导那样的规律可循。有学者感慨，甲骨断痕的边缘并无一定的规律，而人对信息的敏感是有偏好的，此处敏感别处未必敏感，因而遗漏甚多。

计算机没有直觉，只有数字和概率。与人不同，它可以找到没有规律的边缘信息进行匹配。

张展向我们展示如何用计算机辅助缀合甲骨碎片。首先准备一片待缀合的甲骨碎片拓片图像，分辨率精度400dpi（经插值运算获得600dpi），修理甲骨轮廓周围的毛刺，提取段痕边缘的一条曲线。将边缘曲线旋转正负20°，得到同一条曲线不同倾斜角度的集合。用边缘曲线集合与选定的一批甲骨拓片的轮廓线相拟合。在边缘曲线上分多个小段进行采样，计算源甲骨碎片图像与目标甲骨碎片图像边缘采样点之间的距离和，作为不相似度处理。当不相似度值小于某一设定值的时候，意味着可能产生一组成功的缀合。

解释起来有些费劲，但计算机只在瞬息间就能输出缀合结果。

最初，张展跑完程序，得到一组缀合结果，发朋友圈，大家喜出望外。随即，他们得知这组甲骨碎片已被前人缀合过，不过至少证明这个方法行得通。不久后，实验室终于得到新的“独家”缀合结果。不仅文辞能够连上，贯穿两片甲骨的刻痕也明显能够贯通。随着项目深入，得到一组又一组缀合结果。一篇篇对缀合后连接起来的甲骨文句的考释文章接踵发表。

这种缀合方法取得了小小的成功。但要再进一步，还有难关。目前的技术能够让选定的一片甲骨匹配另一片或者一批甲骨。如果要让计算机在大批量甲骨图像中一次找出可缀合的一对或多对甲骨，需要新的算法和更强大的算力。除了技术因素，还有一个问题困扰着研究者。全世界现存约16万片甲骨，分散在15个国家、181家馆藏机构。相比之下，经过整理可供研究且公开发布的甲骨拓片资料就很少了。而机构与机构之间、国家与国家之间的交流合作、资源共享并非易事。

为此，安阳师院团队今年正式启动“全球甲骨数字回归计划”，争取国家、省、市三级政府部门和社会各界的支持，希望到国内外保存甲骨的馆藏机构进行数据采集，让散落各地的甲骨“回家”。这是一个宏伟而又浪漫的计划。凡是用浪漫来形容的事，往往都是很难的，可能需要很多年才能完成。刘永革对张展说：“你看，我搞了一辈子数据库。你一辈子做好甲骨碎片缀合这件事，也就成了。”与数千年的甲骨文相比，人生仿若沧海之一粟。很多事情的成功有漫长的路要走，其待后人乎！