

以多模态大模型推动中国古文字研究发展

莫伯峰¹, 张重生²

(1. 首都师范大学 甲骨文研究中心 北京 100048;

2. 河南大学 计算机与工程学院 河南 郑州 475001)

摘要 人工智能技术与古文字研究的结合已经产生了一批有影响的成果,但几乎都是基于单模态模型。单一模态的智能模型在面对复杂的古文字问题时还存在很大的限制,只能在某些方面起到“辅助”作用,难以独立而完整地解决各种实际问题。文字的形、音、义与各种模态存在不同的关联,汉字的表意属性决定了图像模态和文本模态具有同样重要的作用,中国古文字研究与人工智能的结合须走多模态之路。近年来多模态大模型正在飞速发展,以此为契机来综合性地解决古文字检测、识别、复原等问题,有望取得更好的效果,而文字考释等一些过去难以触及的关键问题也能探索出新的解决路径。

关键词 多模态;人工智能;古文字;甲骨文

Promoting the Development of Ancient Chinese Character Research with Multimodal Large Models

Mo Bofeng, Zhang Chongsheng

Abstract The combination of artificial intelligence technology and ancient Chinese characters research has produced several influential results, but almost all of them are single-modal. The single-modal intelligence model has great limitations in facing the complex problems of ancient characters, and it can only play an “auxiliary” role in some aspects, and it is difficult to solve various practical problems independently and completely. The shape, sound and meaning of characters have different associations with various modalities, and the ideographic properties of Chinese characters determine that the image modality and the text modality have the same important roles, so it is more important to take the multimodal path for the combination of the research of ancient Chinese characters and artificial intelligence. In recent years, the multimodal large model is developing rapidly, which is an opportunity to comprehensively solve the problems of recognition, detection and restoration of ancient Chinese characters, which is expected to achieve better results, and some key problems such as textual interpretation, which were difficult to reach in the past, can also be explored to find a new solution path.

Keywords multimodal; artificial intelligence; ancient Chinese characters; oracle scripts

作者简介:莫伯峰,男,首都师范大学甲骨文研究中心教授,博士生导师,主要研究方向为甲骨学、古文字与人工智能。电子邮箱:mbf2001@163.com。张重生,男,河南大学计算机与工程学院教授,博士生导师,主要研究方向为人工智能、数字人文。电子邮箱:cszhang@henu.edu.cn。

国家社科基金项目“利用神经网络进行甲骨卜辞字体分类的初步研究”(19BYY171);古文字与中华文明传承发展工程规划项目“散见甲骨文图像数据库”(G1930)。

引言

多模态是人工智能中非常重要的一个研究领域。模态是指不同感官的信息,如视觉模态、听觉模态和触觉模态等,多模态模型指的是涉及两种或两种以上不同模态数据或信息(如文本、图像、音频、视频等)的综合处理、分析和理解的模型。在多模态模型中,通过同时处理和学习多种类型的数据,人工智能模型能够获得更丰富的信息表示,进而提高其预测和决策的准确性,使得模型能够更全面地理解和响应现实世界中的多方面信息。多模态模型不仅能拓宽人工智能的应用范围,也使其在理解和模拟人类多感官交互方面具有无可替代的作用,因此被认为是当今人工智能技术发展“第一个前沿方向”^①。

在人工智能的发展过程中,不同模态发展状况存在一定差异。视觉模型是人工智能较早走向成熟并投入实际应用的一种模型。从2012年开始,由于采用深度卷积神经网络来进行图像识别,并利用了图形处理器(GPU)来进行计算,图像识别方面的工作突飞猛进。到2015年左右,某些视觉模型已经开始超越人类的识别率,比肉眼识别更加精准。而自2022年以来,以ChatGPT为代表的大语言模型飞速发展,文本模态模型取得了很大突破,被视为通向“通用人工智能”的可能途径,多模态成为大模型的主要发展方向,GPT-4突出了图文综合处理能力,而最近公布的GPT-4o则进一步融入了音频、视频模态。由此可见,随着各种模态的模型逐渐成熟起来,多模态一体化处理已经具有了较坚实的基础。

大语言模型的出现和发展引起了语言学科的广泛讨论和思考(杨旭,罗仁地 2024),而多模态大模型的发展则无疑为文字学学科提供了一种新的研究角度。文字是记录语言的符号系统,它与语言有着密切的联系,但作为一种视觉符号,它又与语言存在着一定的差异。在视觉模型迅猛发展的时代,文字就已经与人工智能发生了密切关联,文字检测、文字识别、字体辨别、字体设计等视觉模型早已在智能化发展中崭露头角,并投入实际应用。现在,随着大语言模型的兴起,文字学科无疑具有了双重发展机遇。如何有效地将这种双重机遇结合起来,无疑是文字学研究正在面对的一个新课题。

与现代汉字相比,中国古文字具有自己独特的研究对象、研究方法和研究目标,像文字考释等极其复杂的综合性问题,涉及文字的形、音、义等各方面,需要多模态介入。而且这种问题也一定能够得益于人工智能的全方位介入。就如吴振武先生(2020)所说:“像古文字考释这种道理上只允许存在一种正确结论的学问,利用人工智能技术来辅佐研究,自然再适合不过。”实际上,古文字学家研究这些问题时,所采用的方法就对应着综合的多模态处理方式,“就古文字考释而言,人工智能可以提供的帮助是多方面的。著名古文字学家唐兰先生曾将古文字的考释方法总结为形体对照法、推勘法、偏旁分析法与历史考证法。人工智能中的图像识别、自然语言处理、知识图谱等技术正可与这几种方法相互对应(李春桃 2022)。

基于以上认识,本文就多模态大模型与中国古文字研究的关系进行初步讨论。首先,围绕各种单一模态模型在古文字研究中取得的成果,说明人工智能与古文字研究的各种结合

^① 详参孙凝晖:《人工智能与智能计算的发展》(十四届全国人大常委会专题讲座第十讲讲稿),http://www.npc.gov.cn/npc/c2/c30834/202404/t20240430_436915.html。

具有现实可行性。然后,通过分析这些成果背后仍存在的主要问题,以凸显单一模态模型在处理古文字问题时所暴露的短板,特别是面对中国古文字这种具有很强表意属性的文字时所存在的限制。最后,基于一些古文字问题初步的多模态测试,展望未来的发展,指出一些原有课题将有怎样的新研究模式,同时多模态大模型又能够衍生出哪些新的研究课题。

一、各种模态的丰富成果

多模态的发展是以单一模态的成熟应用为基础的。过去人工智能与古文字研究结合主要是以单一模态为主,视觉模型面向图像模态,语言模型面向文本模态,都取得了一些很有影响力的成果。世界各地有各式各样的古文字材料,很多古文字都在与人工智能结合中获得了新成果。而作为世界上唯一没有中断的自源文字,中国古文字与人工智能有丰富的结合点,近年来也取得了很多成果。下面我们简要介绍国外和国内的一些代表性成果。

(一) 国外的主要成果

1. 受损古文字材料的复原

由于古文字材料大多饱受岁月的洗礼非常脆弱,普遍需要对材料上的文字信息进行复原。这种复原能立即推动研究工作的进步,因此这方面的成果很丰富、影响也很大。包括:(1) 古希腊铭文的修复。2022年3月10日,《Nature》封面刊登了著名人工智能团队DeepMind(2022)在古希腊铭文修复方面的重要成果,历史学家与智能模型合作,将受损文字的修复准确率提高至前所未有的72%(见图1a)。(2) 意大利赫库兰尼姆碳化卷轴上文字提取。2024年2月5日,《Nature》网站头版报道了“维苏威火山挑战赛”的结果,意大利庞贝古城的赫库兰尼姆古卷因火山喷发而被碳化,通过深度神经网络辨识出CT图像中的墨迹笔画,成功复原出第一个单词(见图1b)^①。(3) 开罗藏经卷残篇的缀合。2014年人工智能团队通过综合利用视觉特征和单文本行高度及行间距等物理特征,对开罗藏经卷残篇进行了视觉相似性计算,确定候选缀合对象,并使用边缘匹配等方法缩小候选范围,取得了非常好的效果。^②



(a)



(b)

图 1

2. 文字识别

现代文字识别已经是一项非常成熟的技术,作为这种技术的一种自然延伸,古文字识别

^① First passages of rolled-up Herculaneum scroll revealed [2024-02-05]. <https://www.nature.com/articles/d41586-024-00346-8>.

^② Where is my Other Half? <http://www.cs.tau.ac.il/~wolf/papers/otherhalf.pdf>.

很早就受到计算机专家的高度关注。国外的古文字识别主要针对埃及象形文字(Morris et al. 2013)、玛雅文字(Hu et al. 2015)、古印度文(梵文、泰米尔文等)(Munish et al. 2019; Ajoy et al. 2023; Sukhjinder et al. 2023)等,在人工智能深度学习技术兴起之后,这些识别模型的识别正确率都得到了极大提升,而且正开始将识别范围扩展到模型从未见过的字形(Takuru Ishikawa et al.2024),这对于古文字而言非常重要。

3. 文字书写者辨识

在文字识别的基础上,古文字研究通常都需要进行文字风格的辨识,以便更细致地区别材料的不同性质。人工智能对于风格化图像辨识也具有比较成熟的经验,因此一部分研究开始关注到古文字书写者辨识。比如针对西班牙国家博物馆所藏 Avila 手抄版圣经,人工智能团队开发出了两种尺度的辨识模型:一种是根据单个字符特征进行归类,另一种是根据整页文本特征进行归类,两种方式结合使得鉴别效果大大提升(Tara et al.2010, Claudio et al. 2018)。

4. 古代文献辅助理解

人类语言会随着时间而发生演变,古文字所记录的古代语言必然会与后代的语言之间存在差异。因此,利用人工智能技术辅助古代文献的理解,具有切实的应用价值。在大语言模型时代来临之前,这方面的尝试涉及了古梵文、古希腊文等,主要工作集中于词性标注(Pranaydeep et al.2021)、情感分析(Puneet et al. 2023)等方面,并实现了将古希腊语文献翻译为拉丁文(Tariq et al. 2022)。

(二) 国内的重要成果

第一,受损古文字材料的复原。与国外的古文字材料一样,中国古文字同样面临着繁重的材料复原工作。(1) 甲骨缀合。甲骨材料极易破碎,自甲骨文发现不久,甲骨学界就开始了缀合复原的工作。人工缀合工作至今已经一百多年,经历了三个阶段(黄天树 2010),现在终于开始步入智能缀合的阶段。在人工智能的辅助下,已经有一批缀合通过“人机耦合”的方式被发现,缀合效率大为提高(莫伯峰等 2022;莫伯峰,张展 2021;李霜洁 2024)。(2) 竹简编联。竹简出土后顺序是混乱的,只有将它们按照原有顺序编联起来,才能回复原状,因此竹简编联也是一种材料复原工作。进入 21 世纪以来,出土战国竹简大量增加,依靠人工智能进行竹简编联工作可以为专家提供很好的辅助(莫伯峰,胡韧奋 2023)。

第二,古文字识别。由于汉字的表意属性,任何一个时代的汉字数量都要比表音文字多很多。因此,对于汉字古文字的识别是一项十分艰难的工作。尽管如此,近期以来这方面的工作仍然取得了十分可喜的成果,李春桃团队(2023)所开发的古文字识别模型,“经过测试,模型对测试集中的数据识别准确率达到 80.24%”,这是一个非常了不起的成就。

第三,古代文献标点。古文献标点是汉语文献在整理时一项比较特别的任务,无论是传世文献还是出土文献都面临着这项工作。为解决历史上遗留下来的海量文献的标点问题,相关领域的专家相互合作做出了很多努力,为机器自动标点奠定了良好基础。近年来胡韧奋团队基于 33 亿字语料库构建的古汉语模型,提出了新的序列标注方法,断句 F1 值在诗、词、古文三种文体上分别达到 99%、95%、92%以上,使得这方面工作取得了极大的突破(胡韧奋等 2021)。

第四,古代文献翻译。文言文与白话文有着很大的区别,无论是传世文献还是出土文献,最后都需要解决文白翻译的问题。北京师范大学汉字汉语研究与社会应用实验室和数字人文系近期构建了专门适用于古汉语文本理解的大语言模型,能够支持字词释义、文白翻译、用典分析等多种具有挑战性的文言文理解任务。^①这种模型不论是面向普通大众,还是面向专业研究人员,都有极大的价值。

这里只是选取介绍了一些具有代表性的成果。无论是国内还是国外,成果数量正越来越丰富,这是一个非常明显的趋势,相信今后还将有更多的成果涌现出来。

二、亟待突破单一模态的限制

虽然中外成果已经非常丰富,但是这种结合工作现在还处于初始阶段。随着研究的深入,一些困难正在浮现,单一模态限制应该是最主要的瓶颈。下面我们将先分析中外古文字在与人工智能结合时所受限制的一些表现,然后突出中国古文字特性所带来的特有问题。

(一) 中外古文字与人工智能结合在模态方面存在的限制

从上一节的介绍中可以看出,中外古文字研究与人工智能结合都取得了很多成绩。但同时也能看到,在涉及模态问题方面,中外古文字研究面临一些共同的问题:

第一,单模态使得人工智能无法独立而完整地解决问题。在上述研究成果中,常见到的一个词就是“辅助”,以及与专家的分工“合作”,人工智能还缺乏独立完成一个完整任务的能力。其原因在于这些工作通常都是以单一模态介入,所以总会有其他方面能力的缺失。以中外都非常关注的破损古文字材料复原为例,开罗藏经卷、赫库兰尼姆纸莎草纸卷轴的复原是通过视觉模型完成的,在得到结果后需要再依靠人的语言模态来进行验证。而中国古文字中,甲骨缀合任务的完成,是由视觉模型基于甲骨边缘图像相似性进行筛选,然后还需要甲骨学家基于自身甲骨语言能力,在这些备选中确定正确的缀合。

第二,语言模型落后于图像模型,两种模态的发展并不均衡。以上成果中,图像模态的成果占了主体,这与图像类智能模型发展较早、技术较成熟有关。而语言类模型是近期才兴起的一种模态,尽管大模型技术在现代语言任务方面率先取得突破,但是在面对古代语言时还面临一些特殊的困难。而古文字材料作为古代文献中最为特殊的一种语料,最大限度保留了旧时文献的原貌,但也突出了原始文献各种复杂的问题。比如,中国古文字材料中常见的假借现象,一些假借用字的习惯,就与传世文献有较大差异,普通的古代汉语模型难以应对。因此可以说,面向古文字材料的语言模型发展更为艰难。

第三,古文字材料本身还难以支撑大语言模型,需要依靠传世文献建立的语言模型来提供有效辅助。大语言模型的重要基础是大量语料,大部分古文字材料作为一种独立的语料还不够丰富。但是从本质上来说,古文字材料是一种出土文献,它与传世文献在语言性质上并没有本质区别。现代语言与古代语言是一脉相承的,出土文献与传世文献性质相同之处要大于差异之处。古文字学家过去能够取得非常丰硕的成果,也正是基于这种关联性。因

^① 详参黄国学:《AI 太炎:以智能技术 探文言奥妙》, <https://mp.weixin.qq.com/s/Cp5NOSocjvBt9qzcVZ9igQ>。

此,现在面向古文字任务的大语言模型,还必须以传世文献为语料基础进行预训练,再迁移到出土文献上,这是一种具有一定可行性的方法。

(二) 中国古文字与人工智能结合在模态方面存在的限制

在上述共通的问题中,中国古文字研究受到模态单一的影响尤其大。通过比较以上中外成果就可以发现,二者在智能模型所侧重的模态方面存在着比较大的差异。这种差异本质上不是中外差别,而是文字性质的差异、表音文字与表意文字的差异。文字的形、音、义与各种模态存在不同的关联,不同性质的古文字对于不同模态模型的需求状况是不同的,只有从性质差异上理解这种差异,才能知道中国古文字研究需要发展什么样的人工智能模型。

表音文字的基本文字单位是字母,其数量有限,呈严格的线性排布,形体与文字的意义没有关联。因此,这种类型的文字与图形模态关系不是非常紧密,更侧重于语言模型的应用。表意文字的基本文字单位是偏旁部首,由偏旁部首构成单字。偏旁部首通常可达数百,并且相互之间不呈现严格的线性排布,单字数量则更为庞大,可至数千甚至上万,这也就意味着有数千甚至数万个不同的图像符号需要识别,这是一项非常艰巨的任务,因此图像模态不可或缺。与此同时,文字是对语言的记录,表意文字背后反映的仍是一种语言结构,表意文字对于语言模型的需求并不比表音文字低。

与现代汉字相比,中国古文字的表意属性更加显著。裘锡圭先生(2007)指出:“汉字在象形程度较高的早期阶段(大体上可以说是西周以前),基本上是使用意符和音符(严格说应该称为借音符)的一种文字体系;后来随着字形和语音、字义等方面的变化逐渐演变成为使用意符(主要是义符)、音符和记号的一种文字体系(隶书的形成可以看作这种演变完成的标志)。”因此,中国的文字学研究有一些其他文字没有的研究内容,比如从许慎《说文解字》开始,就关注文字的“本义”,也就是汉字字形所体现出来的字义。这对于不具有表意性质的文字而言,显然是没有必要的。中国的古文字学家在解决具体问题时,需要综合考虑形、音、义各方面的情况,如著名古文字学家杨树达(1997)所说:“每释一器,首求字形之无牾,终期文义之大安,初因字以求义,继复因义而定字。义有不合,则活用其字形,借助于文法,乞灵于声韵,以假读通之。”这种综合性的方法,对于人工智能而言就是一种多模态的思想。由此可见,中国古文字的形、音、义都需要有专门的智能模型,而最终还需要把这些不同模态的模型统一起来,才能解决真实的研究性问题。

由于表音文字和表意文字的这种性质差异,导致了二者与人工智能不同模态结合时呈现出了一些非常不同的特征。通过比较一些共同的研究题目,就能看到中国古文字更需要多模态才能解决这些问题:

第一,古文字材料复原的方法。表音文字通过在不同模态间进行转换,可以更少依赖多模态。大多数古文字材料复原,都是一个图像问题。但对于表音文字而言,则还可以将其转化为语言问题来处理。比如前述 DeepMind 复原古希腊文字的模型就是一个语言模型,它通过把图像的问题转化为文本的问题来加以处理,而将图像方面问题交给专家,因此“与他们合作的历史专家在单独修复古代文本时准确率只有 25%,但当他们与模型合作修复时,准确率提高到了 72%”(Yannis et al. 2022)。而这种方法如果放到中国古文字领域则存在很大的困难,以甲骨的缀合为例,如果把甲骨上的内容转录为文本,完全以文本为根据进行缀合复

原,就不具有可行性。因为甲骨文字除了记录语言,自身形体还蕴含着很多其他视觉信息,比如记录同一词语的异体字字形,对于缀合就有非常重要的意义,舍弃了这些图像信息进行复原困难会非常大。

第二,对文字识别模型的需求。表音文字由于字符数量有限,不会特别需要专门的文字识别模型。而对于中国古文字而言,历史上积累下来的文字数量十分庞大,这些文字的风格又有很大的差异,不仅识别文字具有极大的挑战性,就是进行文字的检测,即定位哪里有文字,也是一项很困难的任务。而且,由于汉字是一脉相承至今没有中断的一种文字,古汉字识别本质是先将手写古文字识别为有编码的古文字,然后还要将古文字与现代文字进行对应。这与国外很多的古文字没有后世文字的延续,识别类似于一种语言翻译的情况完全不同。因此,文字识别模型对于中国古文字而言绝不是可有可无的,而是有非常大的应用空间,在面对各种识别中的困难时,还需要结合语言模型来综合解决。

第三,辨识文字书写者的难度。这个问题同样受到了文字性质差异的影响,区别文字书写者最小的尺度是字符,对于表音文字而言,在同一作品中很容易找到一定数量的相同字母进行模型训练,而在不同作品之间也较容易找到相同字符进行比较。而对于中国古文字而言,这就存在一定的困难,很多文字出现的次数很少,就难以据此进行辨识。表意文字不仅字符数量大导致不容易重复出现,而且表意文字存在常用字和罕用字的分布,也就是字符使用频率的差别,这就更加使得对相同字符进行比较变得困难。因此,虽然通过相关的测试可以发现技术路线上是可行的(莫伯峰 2019),但是在应用时常常会因相同文字不充足,而难以发挥作用。在实际的文字书写者辨识工作中,只有结合语言模型,从语言中发现差异,才能辅助这个问题更好地得以解决。

第四,综合性任务的问题复杂程度。文字考释是中国古文字学的一个核心研究内容,而表音文字基本上就不存在这项任务。“考释文字”并不等同于解读某个文字所包含的词义,还要找到古文字对应于后代的哪个字。汉字是世界上唯一从古到今一直运用、没有中断的文字体系,所以这种在早期字形和晚期字形之间建立联系的任务才会存在。由于中国古文字的字形存在相当多的变化,使得这种对应工作变得极其复杂。如果忽略了图像方面的因素,也就舍弃了最为重要的部分。因此,对于中国古文字而言,这种综合性任务的复杂程度远高于表音文字,必须利用多模态大模型才能全面地解决。

综上所述,汉字的表意属性决定了图像模态和文本模态具有同样重要的作用,中国古文字研究与人工智能的结合更需走多模态道路。从这些既有成果来看,中国古文字的性质也决定了只有突破单一模态的限制,才能取得更为长远的发展。

三、多模态大模型为中国古文字研究开拓更广的道路

从人工智能的演进来看,多模态是一个明确的发展方向。从中国古文字的性质来看,多模态大模型是推动研究深入发展的必然要求,这方面的探索值得积极尝试。由于现在还缺乏专门面向中国古文字的多模态模型,我们基于 GPT-4o 和 Claude 3.5 Sonnet 这两种通用领域的多模态模型进行了一些初步测试,通过模拟人类古文字学家在考释古文字时所使用的三种基本方法,来初步评估这种发展思路的可行性和主要障碍。

1. 偏旁分析法




“对不认识的字,通过分析,确定构成它的各个偏旁,将这些偏旁与已识的字相比较,再组合起来认识所要考释的字,这种方法就是偏旁分析法”(黄德宽 2019)。偏旁分析法是考释古文字的一种基本方法,对表意文字而言,是辨析字形了解字义的一种重要手段。




在测试中,我们向 GPT-4o 和 Claude 3.5 Sonnet 输入了图片形式的文字,并用提示语的形式输入我们的要求。由于 GPT-4o 和 Claude 3.5 Sonnet 的回复普遍较长,限于篇幅这里用表 1 简要介绍一下测试的内容以及所得的基本结果,并在表后做一些说明(见表 1)。

表 1

| 输入图像 | 特 | 𠂔 |  |  |
|-------------------|------|------|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| 输入要求 | 文字切分 | 文字切分 | 切分后预测现代汉字 | 切分后预测现代汉字 |
| GPT-4o | ✓ | ×、✓ | ✓ | ×—✓ |
| Claude 3.5 Sonnet | ✓ | ×、✓ | ✓ | ×—✓ |

- (1) 以上测试文字按照由易至难的标准选取。
- (2) “𠂔”为生造汉字,以避免模型对文字的切分知识来自语料中的信息。
- (3) 通过测试可知,模型可以根据要求,对文字进行多种层次切分。比如“𠂔”字可以切分为“客”和“辱”,也可以切分为“宀”“各”和“辱”。无专门要求时,模型倾向于切分为三个偏旁。也可以要求模型将所有的切分层次都罗列出来,但切分的效果很不稳定,“𠂔”右边的“辱”基本无法切分正确,GPT-4o 只切分为一个偏旁“辰”,Claude 3.5 Sonnet 切分为“辰”和“尸”。同时,GPT-4o 和 Claude 3.5 Sonnet 虽然也具有文生图功能,但还完全不能根据对偏旁的描述来生成相应的汉字图像。

(4) 在提示语中告知模型“”为甲骨文“人”字,让模型预测甲骨文“”(众)字。虽然甲骨文“”字比“众”字多了上部的偏旁“日”,但因为不容易受类似字形影响,仍然预测正确。

(5) 在提示语中告知模型“”为甲骨文“从”字,“”为甲骨文“林”字,让模型预测甲骨文“”(休)字。GPT-4o 和 Claude 3.5 Sonnet 第一次都回答错误。补充提示“所预测甲骨文由左右两个偏旁组成,请重新作答”后,预测正确。

通过这个简单测试可以看到,GPT-4o 和 Claude 3.5 Sonnet 在面对汉字时,确实展现了一定的多模态功能,能够同时处理图像和文本两种信息,而且有一定的正确率,但所存在的问题是预测效果不稳定,对文字的切分有时会出现错误,且不具有文字图像的输出能力。

2. 字形比较法

“所谓字形比较法,具体说来就是利用汉字系统性和古今发展的相互关系,拿已经认识的字(或偏旁)与未识字(或偏旁)做形体上的细致对比来考释未识字。”(黄德宽 2019)字形比较法也是释读古文字的一种基本方法,它要求从大量的字形演变中总结出变化规律,相较于偏旁分析法需要更强的图像分析能力。

在测试中,我们向 GPT-4o 和 Claude 3.5 Sonnet 输入了“牛、羊、马、豕、犬”五个汉字沿

着甲骨文、金文、战国文字、篆文、隶书、楷书的时间演进顺序不同的字形。要求模型结合汉字的这种演进规律,预测新见到甲骨文是哪个现代汉字。

我们测试了甲骨文“𩺰”(鱼)、𩺰(佳)、“𩺰”(龙)、“𩺰”(王)等较为简单的独体字。GPT-4o 和 Claude 3.5 Sonnet 都只预测正确了“鱼”字,但其预测的方式并非从字形演变规律中来。而是来自对文字图画性的分析,两个模型强调了“类似于鱼的鳍和尾巴”“明显呈现出鱼的轮廓,包括头部、身体和尾部”这种因素,因此“推测很可能是‘鱼’字的早期形态”。

因为所输入的文字演变例字较少,而且缺乏中间过渡字形的连接,因此模型肯定难以学习到文字演进中复杂的规律,要通过“字形比较法”来预测古文字,还需要完整而系统地导入汉字演变的图像数据,并且开发相关的专门模型。GPT-4o 和 Claude 3.5 Sonnet 这种通用型的大模型显然还难以应对这样的任务。

3. 辞例归纳法

“辞例归纳法,是依据未识字出现的语言环境,通过对一系列辞例的分析、比较、归纳,从而达到释字目的的方法。”(黄德宽 2019)这种方法强调了考释古文字需要语言模态发挥作用,但是辞例只能卡定一个范围,所以还需要结合字形进一步缩小范围。我们进行了两种类型的测试,分别预测词义和文字。

第一种是基于甲骨文“众”字的识别结果,让模型根据语境预测“众”的词义。提示语为:上述图片中的甲骨文“众”字出现在这样一句商代卜辞中:“丁亥贞:王令众伐召方,其受祐?”请分析在这句话中,“众”表示什么意义?

GPT-4o 回复:商代卜辞“丁亥贞:王令众伐召方,其受祐?”中,“众”字的意义可以通过上下文和甲骨文的象形特点来分析。上下文分析:这句话的意思是:“在丁亥日占卜:王命令众人去讨伐召方,他们会得到神灵的保佑吗?”其中,“王令众伐召方”表示王命令一群人去讨伐召方。“众”字的象形意义:甲骨文中的“众”字由三个“人”字组成,表示许多人的集合。在这句话中,“众”字指的是一群人或军队。具体意义:在这句话中,“众”具体指的是王所命令的那群人或军队,强调了人数的众多。这群人是执行讨伐任务的主体。综上所述,在这句商代卜辞中,“众”字表示一群人或军队,强调了人数的众多,是王命令去讨伐召方的主体。

第二种是基于字形和语境,预测文字。在测试中,我们向模型 GPT-4o 和 Claude 3.5 Sonnet 输入了甲骨文“𠂔”(从)字和“𠂔”(步)字的图像。并用提示语要求模型预测甲骨文“𠂔”(先)字。

GPT-4o 回复:最有可能是“攸”字。理由如下:“从”字由两个“人”字组成。“步”字由两个“止”字组成。由左边的“人”和右边的“止”可组成“攸”字。

Claude 3.5 Sonnet 回复:“从”字由两个并列的人形组成。“步”字由两个“止”字组成。所要预测的甲骨文看起来是由一个人形和一个“止”字组合而成,我们推测最可能是:“徒”字。

可以看到,由于不具有对“先”字偏旁部首的认知,两个模型都不能通过字形正确预测出“先”字。我们通过增加语境,从语言的角度来对这个字进行预测。提示语为:这个字出现的语境有(句子中的 A 代表这个字):1. 庚子卜殷贞:勿令子商 A 涉羌于河。2. 甲申卜争贞:勿乎妇好 A 収人于靡。3. 丁卯卜争贞:令子效 A 于甸。4. 王令望乘 A 归。重新预测这个甲骨

文是什么字?^①

GPT-4o 回复:根据提供的甲骨文和语境信息,这个甲骨文最有可能是“徙”字。理由如下:甲骨文“从”字和“步”字都与行走、移动有关。语境中的句子提到“勿令子商 A 涉羌于河”、“勿乎妇好 A 収人于靡”、“令子效 A 于甸”、“王令望乘 A 归”,这些句子都涉及移动、迁徙、行走等含义。所要预测的甲骨文由左右两个偏旁组成,左边是“从”,右边是“止”,符合“徙”字的构造。因此,结合甲骨文的形状和语境,最有可能是“徙”字。

Claude 3.5 Sonnet 回复:根据这些新的线索,我现在可以更准确地推测这个甲骨文字符。考虑到您提供的语境和甲骨文的形状,这个字很可能是“使”字的甲骨文形式。

显然模型将这个字预测为一个动词,我们继续补充提示语:从语境来看这是一个副词,而不能是一个动词,再重新进行预测。两个模型都推测文字为“迺”,离正确答案更远。

从以上的结果能够看出来,现在的多模态模型距离能够发挥实际研究作用还有非常远的距离。但是,多模态模型展现出的这些功能必然是古文字研究需要用到的。在人工智能发展日新月异的这个时代,相信随着这些功能逐渐完善,古文字识别、检测、分类、复原等问题都能得到更好的解决,而文字考释这种古文研究的核心任务,也必将随着多模态大模型的发展获得更多的发展契机。

习近平总书记在致信祝贺甲骨文发现和研究 120 周年时指出:“新中国成立 70 年来,党和国家高度重视以甲骨文为代表的中华优秀传统文化传承和发展,多部门多学科协同开展甲骨文研究和应用,培养了一批跨学科人才。经过几代人辛勤努力,甲骨文研究取得显著成就。”“多学科协同”一直是中国古文字学得以持续发展的动力,在人工智能飞速发展的今天,不断探索人工智能与古文字研究结合的方式,以问题意识引领未来的研究模式,这不仅是中国古文字学科进步的必然要求,也是交叉研究发展的必由之路。中国古文字学本身就具有很强的交叉学科性质,与考古学、历史学、语言学、文献学等都有密切的联系,由此也决定了其多维的研究角度(李学勤 2006)。在中国古文字学发展过程中,要使得这些多维度都能够推进学科发展,依靠人工智能的多模态大模型是一个重要的路径。

参考文献

- 胡韧奋,李 绅,诸雨辰 2021 《基于深层语言模型的古汉语知识表示及自动断句研究》,《中文信息学报》第 4 期。
- 黄德宽 2019 《古文字学》,上海:上海古籍出版社。
- 黄天树 2010 《甲骨拼合集·序》,北京:学苑出版社。
- 李春桃 2022 《人工智能如何辅助古文字研究》,《光明日报》10 月 30 日第 5 版。
- 李春桃 2023 《基于人工智能技术的古文字研究》,《吉林大学社会科学学报》第 2 期。
- 李霜洁 2024 《人工智能引导人类直觉产生的甲骨新缀第 21—30 组》,《出土文献》第 2 期。
- 李学勤 2006 《古文字学初阶》,北京:中华书局。
- 莫伯峰 2019 《利用深度神经网络判别甲骨文字体的初步测试——以字形为依据》,《甲骨纪实》,北京:故宫博物院出版社。
- 莫伯峰,胡韧奋 2023 《利用人工智能技术进行竹简编联的初步测试——以〈清华简〉为例》,《古文字与出

^① 所使用的辞例皆来自真实的甲骨文卜辞。

- 士文献青年学者论坛(2019)论文集》,上海:上海古籍出版社。
- 莫伯峰,张重生,门 艺 2022 《AI 缀合中的人机耦合》,《出土文献》第 1 期。
- 莫伯峰,张 展 2021 《计算机辅助缀合研讨一则》,《民俗典籍文字研究》第 28 辑。
- 裘锡圭 2007 《文字学概要》,北京:商务印书馆,2007 年。
- 吴振武 2020 《古文字考释与人工智能》,《光明日报》11 月 7 日第 12 版。
- 杨树达 1997 《积微居金文说》,北京:中华书局。
- 杨 旭,罗仁地 2024 《ChatGPT 来了:语言科学如何看待 ChatGPT》,上海:上海教育出版社。
- Ajoy, M. & C. V. Jawahar. 2023. Competition on Indic handwriting text recognition//2023 International Conference on Document Analysis and Recognition, Switzerland: Springer Nature, 435 - 453.
- Claudio, De T., M. Maniaci, F. Fontanella, et al. 2018. Reliable writer identification in medieval manuscripts through page layout features: The “Avila” Bible case. *Engineering Applications of Artificial Intelligence*, 72.
- Kavitha, B. R. & C. B. Srimathi. 2022. Benchmarking on offline Handwritten Tamil Character Recognition using convolutional neural networks. *Journal of King Saud University-Computer and Information Sciences* 34(4), 1183 - 1190.
- Morris, F., C. Jan & van Gemert. 2013. Automatic Egyptian hieroglyph recognition by retrieving images as texts//The 21th ACM International Conference on Multimedia, Spain: Barcelona, 765 - 768.
- Munish, K., M. K. Jindal, R. K. Sharma, et al. 2019. Character and numeral recognition for non-Indic and Indic scripts: A survey. *Artificial Intelligence Review* 52(4).
- Pranaydeep, S., G. Rutten & E. Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek//5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, co-located with EMNLP 2021, Dominican Republic: Punta Cana, 128 - 137.
- Puneet, K., K. Pathania & B. Raman. 2023. Zeroshot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data. *Applied Intelligence* 53(9), 10096 - 10113.
- Hu, R., G. Can, C. P. Gayol, et al. 2015. Multimedia analysis and access of ancient Maya epigraphy: Tools to support scholars on Maya hieroglyphics. *IEEE Signal Processing Magazine* 32(4), 75 - 84.
- Sukhjinder, S., N. K. Garg & M. Kumar. 2023. Feature extraction and classification techniques for handwritten Devanagari text recognition: A survey. *Multimedia Tools and Applications* 82(1), 747 - 775.
- Takuru, I., T. Miyazaki & S. Omachi. 2024. Japanese historical character recognition by focusing on character parts. *Pattern Recognition*, 148.
- Tara, G., R. C. Wilson & J. A. Clark. 2010. Scribe identification in medieval English manuscripts//2010 20th International Conference on Pattern Recognition(ICPR 2010). Turkey: Istanbul, 1880 - 1883.
- Tarrq, Y., C. Palladino, D. J. Wright. 2022. Automatic translation alignment for ancient Greek and Latin//Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, France: Marseille, 101 - 107.
- Yannis, A., T. Sommerschield, B. Shillingford, et al. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature* 603(7900), 280 - 283.