基于人工智能技术的古文字研究

李春桃　张　骞　徐　昊　高嘉英

【摘　　要】人工智能与古文字学交叉研究十分重要，开展这项研究既需要人工收集和标注大量数据，同时也需结合恰当的技术。在数据处理方面，数据集建设过程中尽量丰富了单字数量以及字图总量。数据中的字图包括拓本和摹本，其中拓本多带有斑点噪声，降低噪声有助于提高文字识别的准确率。数据中古文字隶定体的显示也是要重点解决的问题。在文字自动识别方面，利用了深度学习算法开展智能识别，从实验结果看，准确率达到八成以上，这是在大规模识别任务下达到的效果，证明了利用人工智能技术识别古文字形体是可行的。分析错误数据可以发现，数据量与形近字是影响识别准确率的关键因素。除了识别以外，知识图谱技术也很重要，建设古文字知识图谱一方面可以实现对古文字知识体系的多角度展示；另一方面也可计算字形中偏旁及构形的相似度，智能寻找出字形之间的联系。

【关 键 词】人工智能；古文字研究；深度学习；知识图谱

【作者简介】李春桃，吉林大学考古学院、古籍整理研究所教授，历史学博士；张骞，吉林大学考古学院博士研究生；徐昊，吉林大学计算机科学与技术学院教授，工学博士；高嘉英，吉林大学人工智能学院博士研究生(长春 130012)。

【原文出处】《吉林大学社会科学学报》(长春)，2023.2.164～173

【基金项目】国家八部委“古文字与中华文明传承发展工程”资助项目(G3829)；国家社会科学基金项目(18BYY135)。

近些年来人工智能发展迅速，尤其是深度学习技术，具有学习知识、分析问题、总结规律的能力，能够对文字、图像和声音等数据进行识别、归纳与分类。鉴于此，已有学者将深度学习应用于汉字的识别任务中，尤其是对手写汉字的识别取得了不错的效果。这也提示我们将人工智能运用于古文字形体识别是可行的。相比于偏重主观感受的学科，古文字研究更为客观，其结论具有唯一性，研究过程也遵循一定的规律，在人文学科中最接近自然科学。这些都与人工智能技术的工作原理相互契合。

已有学者在这方面进行了探索，但更多的是技术性的尝试，或是理论上的思考，尚缺乏系统性的大规模研究。这可能与古文字材料自身的特点有关。首先，古文字与现代文字存在很大区别，在数据处理以及技术结合上都需要大量的专业知识，而掌握古文字专业知识的学者属于小众群体，并不具有普遍性。其次，人工智能研究需要高质量的数据集，目前来看形体数量庞大且单字丰富的公开数据集几乎没有，需要单独构建。再次，由于出土资料有限，古文字形体数量多寡不一。有的常用字可能出现数千次，字图数量也能达到数千个；而有的文字仅出现一两次，且字图数量也仅有一两个，后一种情况在古文字中占比很大。在数据不足的情况下，人工智能模型难以学到泛化的分类特征，会对识别准确率产生较大的影响。最后，古文字形体的呈现方式主要是拓本，很多拓本上存在腐蚀噪声，会对模型提取字图的特征形成干扰，而对拓本进行降噪本身也是一个复杂的问题。以上因素都在不同程度上影响了古文字智能化识别的进程，是古文字与人工智能交叉研究领域需要面对和解决的重要问题。近几年来，我们在这一领域做了一些探索，收集整理了古文字数据，并对数据进行了

分类与标注，利用深度学习算法完成了识别实验，同时也就古文字知识图谱的构建做了初步尝试。

一、数据的收集与处理

(一)数据的收集与增强

古文字形体以拓本为主，而一些特殊资料又以摹本形式呈现[①]，所以在建立数据集时我们根据这种实际情况，既收集了大量拓本，也利用了已有的部分摹本，其中拓本占绝大多数。数据集中包括甲骨文、金文以及战国文字，其中战国文字包括印文、陶文、币文，而竹简文字则以早年发表的为主，近年新公布的清华简、安大简等材料尚未收录。在数量方面，第一次完成的数据集中字图总量为150680张。随着不断扩充，近期又更新了数据集，最新一版的字图总量是556390张，以甲骨文、金文为主。[②]在单字数量方面，以往的研究所覆盖的范围都不够丰富。为了确保研发数据的充分和全面、得到的实验结果更加客观，我们在收集过程中有意增加了单字数量，数据集中单字达6941个。与以往研究相比，单字数量是最多的。

我们收集的数据总量颇为丰富，却呈现出不平衡的特点。有的文字图版数量庞大，如“亡”“年”“田”等字在古文字中出现了数千次，尽管我们未将其字图全部收录，但在数据集中每个字也高达一千余个；而那些仅出现一至两次的形体，虽已全部收入，但其样本总量仍十分匮乏。这使数据集在结构上分布不平衡，使用不平衡的数据，模型在学习过程中容易导致特征偏移。为了解决上述问题，我们进行了“数据重构”，对于样本数量超过阈值上限的数据采取随机采样方法，即对数据集中某些单字存在大量重复、冗余的样本进行随机抽样，可以简化样本空间中重复的特征点，降低计算复杂度，同时也可在一定程度上降低训练过程中出现的过拟合效应。对于样本数量低于阈值上限的数据采取数据扩充的方法，利用计算机图形算法将图像进行不同程度的变换，包括仿射、剪裁、调色以及旋转等方式(参图1)，进而实现数据量的增加。通过数据的扩充，智能模型可学习到更多的分类特征，也提高了泛化能力。

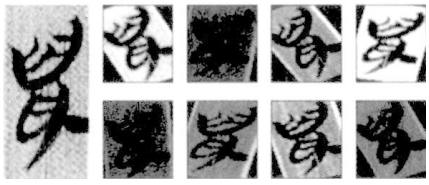

图1 数据扩充

(二)字体的生成与显示

古文字中有很多字不见于今天的楷书，由于现在通用的字体库中不包含这些字符编码，所以这些文字形体可以隶定，却无法以字体形式显示。如  (《合集》7287臼)形从女从羊[③]，可隶定作“姅”；  (《铭图》15155)形从亦从示[④]，可隶定作“祘”；(《玺汇》4039)形从门从犬[⑤]，可隶定作“閔”，以上形体都可进行字形拆分并加以隶定，但这些隶定字无法以字体形式显示。古文字学者在撰写文章涉及此类问题时，往往以图片形式来显示隶定字，但人工智能模型输出时则不宜采用图片形式。故而，欲实现大规模的古文字识别，需先解决古文字字体显示问题。在字体方面，中华字库工程曾经研发了输入法以及“中间字库”，可以提供很多便利，但古文字中还有一些字形没有包含在“中间字库”之内，为此我们研发了“集大字库”，把那些可以隶定但无法显示的字形收入其中，如此处理之后这些隶定字便可在模型中输出。除此之外，“集大字库”对于我们为古文字材料所作的

释文资料也有很大帮助，不仅使这些字形能够以隶定字显示出来，还可以实现检索功能。

(三)拓本的降噪与处理

甲骨文与金文的主要呈现方式是拓本，而由于文字载体本身的残损、锈蚀等原因，拓本往往会出现一些斑点、泐痕等非文字笔画痕迹，图像中这些非必要的或多余的干扰信息在计算机领域被称作噪声。带有噪声的拓本如  、  、  、  、  ，类似情况一般不会给古文字研究者造成影响，因为专家凭借知识积累以及研究经验，很容易排除这些噪声，但是对于人工智能模型而言，噪声会形成较大的障碍，所以在人工智能与古文字的交叉研究中，降低噪声是很重要的步骤。以往一些研究文字识别的学者，较多利用的是摹本而非拓本，最主要的原因就是拓本存在噪声。对拓本图像进行降噪处理是十分必要的工作。

我们先后采用腐蚀化、骨架化、膨胀化、二值化的方法，最终实现了图像降噪的目标。例如伯椃虘簋铭文中的“皇”字作  (《铭图》5085)，该形左侧和右上部都有噪声。在降噪过程中(参图2)，首先对其进行“腐蚀化”操作[6]，尽量减少拓本中的小面积独立噪声，当然这一操作会使文字笔画受到部分影响；接着采取“骨架化”操作，提取拓本中文字的形体骨干，噪声多数会在前两个步骤中被排除；然后进行“膨胀化”操作，将形体骨干加粗，重新变成丰满的笔画；最后是“二值化”操作，将拓本处理成白色文字和黑色背景的形式。在实际研发过程中，数据集中每一个文字拓本都会经过这一降噪过程，从而弱化图像中的噪声干扰，提高模型对笔画特征的提取能力，增强模型分类的准确性。

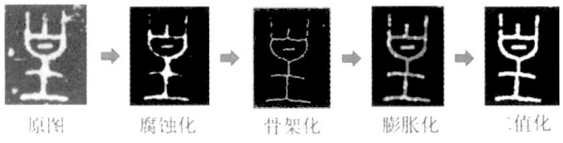

图2 图片降噪

二、古文字智能识别

(一)智能识别的原理与实验

在利用人工智能识别古文字形体时，首先应该了解研究者是如何考释古文字形体的，然后再结合计算机技术设计和研发出适合的方法。关于古文字的考释方法，著名古文字学家唐兰先生曾总结为形体对照法、推勘法、偏旁分析法与历史考证法。[1]163-193这几种方法对于今天的古文字考释仍然适用。当然，近年随着战国竹简的陆续公布，古文字的考释方法也发生了变化，通过破解通假关系找到文字所代表的“词”显得尤为重要。详审两者区别，前者可概括为“考释文字”；后者可概括为“考释词语”。就“考释文字”而言，字形是十分重要的。于省吾先生曾考释出大量的甲骨文，在总结考释经验时，他提到“留存至今的某些古文字的音与义或一时不可确知，然其字形则为确切不移的客观存在，因而字形是我们实事求是地进行研究的唯一基础”[2]3-4。林沄先生曾将其总结为“以形为主”[3]37。这种方法显然是科学的，唐兰先生总结的形体对照法、偏旁分析法与历史考证法都是针对古文字形体而言。

在人工智能视觉领域，主要基于深度学习技术来识别古文字，用足够多的字形图版训练深度学习模型来达到识别目的。这种技术显然也是着眼于字形，是以形体本身为研究对象和出发点，可以类比的是上文谈到的形体对照法。形体对照法是古文字考释中最为直接和常用的方法，是把不同材料中古文字形体加以比较、

对照，利用已识字来考释未释形体。针对古文字的特殊性，我们选用ResNet18基础神经网络模型并对参数进行了调整。该模型具有特定的浅层网络结构，可以针对图形数据进行特征提取，同时具备收敛速度快、识别准确率高、训练成本低等优点。

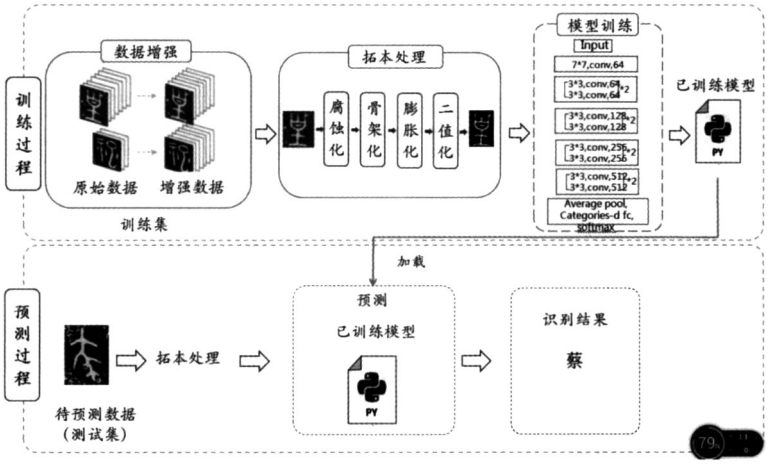

图3 识别过程

整个流程可分为两个步骤(图3)：训练过程和预测过程。在训练过程中，首先会对文字图像进行数据扩充，然后对每一张拓本进行降噪处理，经过处理后，文字图像变得清晰。将处理后的图像输入网络模型进行训练，在训练中模型可以提取字形的关键特征并将图像分类至对应的文字标签下。在迭代训练数十轮之后，我们将模型参数进行保存，得到了已训练模型。

预测过程是检验模型识别能力阶段。将待测试字形进行处理，然后输入至已经被训练过的网络模型中，输出结果。经过测试，模型对测试集中的数据识别准确率达到80.24%。我们使用的是以拓本为主的数据，字形又来源于不同的时代，时间跨度较大，而且是在大规模识别任务下达到的效果，所以这个准确率是比较理想的。

(二)实验结果的总结与分析

测试的准确率达到八成，对这些测试结果进行总结，会给我们以启发；同时，还有两成数据识别错误，分析这些错误的数据也会给我们带来更为深入的思考。

从时代上看，甲骨文与西周金文的识别准确率更高，而战国文字的识别准确率较前两者低，后者较前两者的平均值低了6.34%。这主要因为甲骨文与西周金文写法更为稳定，同一个字在不同的材料上虽略有变化，但总体而言并不剧烈，智能模型能够把这些形体合理地联系起来，给出正确的判断。战国文字则不然，一方面，战国时期同一个字在不同地域写法区别很大，即“文字异形”[4]；另一方面，战国文字出现了大量的草率写法，简体、俗体、异体都大量涌现。这些都给识别工作带来了困难。举例来说，战国时期楚系文字中的“夏”字异体较多，仅《楚系简帛文字编》就收录了七种异体[5]525-527，其常见写法作  、  、，也有很多变体，如在上博简《容成氏》中作  (47号)[6]139，“页”旁被省略，形体与“蜀”字(老子甲21)类写法相近[7]，模型便误释作“蜀”。“夏”在上博简《缁衣》中作  (18号)[8]62，这种写法是在上一种的基础上把“虫”形上部拉得平直一些，形成横画，进而

与甲骨金文中的“昜”字(《铭图》5322)相同,形成了不同时期的“同形字”,所以《缁衣》“夏”字被智能模型误释作“昜”。“夏”字还有一种变体,在包山简中作  (包山225)[9]图版九九,此形左部省略成“日”形,右面省略成“首”形,上下结构,其出现次数极少,我们的数据集中仅有一例,模型误释作“为”。⑦从“夏”字的情况可以看出,战国文字变化剧烈、异体纷繁,导致战国文字的识别准确率低于甲骨文与西周金文。

从字图呈现方式上看,摹本的平均准确率较拓本低4.64%。按照常理,摹本更为清晰,没有噪声干扰且笔画明确,准确率应当更高,这也是过去一些研究者更愿意使用摹本的原因所在。实验结果之所以呈现相反的效果,应当与数据有关。数据方面,我们搜集的字图中拓本占绝大多数,摹本占比仅为15.13%。智能模型在经过反复训练后,对拓本上字形的特征更为敏感,也更易于提取,所以在测试中拓本的表现超过了摹本。

影响识别准确率方面,有两个因素表现得最为突出:一是数据量;二是形近字。前者容易理解,一些古文字形体仅出现一两次,即使做了摹本,在数量上依然不够丰富,导致模型学习的特征过于单一,无法进行准确分类。对于深度学习算法来说,数据的质量以及分布对模型的训练有着巨大影响,这不仅局限于古文字专业,其他领域也是如此,无需赘述。下面重点讨论第二种影响因素。同一时代的文字中便已存在很多形近字或者同形字,如甲骨文中的“月”和“夕”、“士”和“王”、“上”和“二”等。如果时间范围扩大,形近字的数量更多。我们的数据范围涵盖了甲骨文、金文、战国文字,从时代上看,早至商代晚到战国末年,历时千年左右,其间文字的载体、用途、形体都发生了变化。理想的情况应根据时代、性质或者载体对数据进行区分,分别实验⑧,这不仅在收集数据方面更加容易,识别难度也大幅度降低。但是,此次我们的目标是考察智能模型对先秦古文字的整体识别情况,检验其对不同时代文字的系联能力,所以有意将不同时代的文字都放入同一数据集中。当然,这也导致数据中存在大量的形近字甚至是同形字,给模型造成较大的干扰。查验测试数据,在识别错误的形体中,形近字占绝大多数。除了大家都了解的本身写法就相近的文字外(如上文所举“月”与“夕”等),还有很多因为形体讹变而偶然相近的。如“虎”字在伯晨鼎中作  (《铭图》2480),与常见“虎”字写法有别,这是因为伯晨鼎铭文比较特殊⑨,致使该形与“虔”近似,人工智能模型也误释作“虔”。宰甫卣铭文中“光”字作  (《铭图》13303),“光”在古文字中属于常见字,尤其在甲骨金文中,下部都从侧面跪坐人形  ,而古文字中跪坐人形有时可以替换成女形⑩,所以卣铭中“光”字下部可以写成“女”。如此一来便与“每”字近似,智能模型也将其误认作“每”。此类现象是比较多的,如甲骨文中的“子”字  (《合集》21567),被误认成形体相近的“列”[10];金文中的“生”字或作  (《铭图》2392),被误认成形体相近的“之”;楚简中“安”字或作  (《孔子诗论》3号)[8]115,被误认成形体相近的“民”;楚简中“天”字或作  (《民之父母》2)[6]118,被误认成形体相近的“而”;玺印中“强”字或作  (《玺汇》2749),被误认成形体相近的“侃”;玺印中“兵”字或作  (《玺汇》3445),被误认成形体相近的“共”。类似因形体变化而被误识的例子极多。需要注意的是,上举例证,很多属于“个案”。如“光”字写作形是偶然现象,在金文中仅出现两例⑪,数量极少,模型才会将其归类到“每”字中,而其他正常写法的“光”字多会被正确地识别。那些偶然讹变而被误识的形体,主要是由于形体混同导致,也与这些讹体的

数据量不够丰富有关。可见数据量与形近字两个因素彼此之间是相互影响的。

影响识别准确率的还包括人为因素。数据的整理及标注过程如果存在矛盾和误差会直接影响最终的识别结果。我们的数据由团队成员分批次进行收集整理,最终再统一、汇总。由于前后的认识存在差异,标注也会不同。如古文字中有两类"御"字:一类正常写法作  (《铭图》15585);还有一类作  (《铭图》5387),两者形体数量都颇为丰富。《说文》:"御,使马也。从彳、从卸。驭,古文御从又、从马。"[11]37《说文》没有把"驭"字单独设立成字头,而是当成"御"字古文,所以常用的字编类工具书多把类形体收录在"御"字头下。[12]115[13]212我们在处理金文时也采用了同样方式,将这两类形体都收在"御"字下。而后来在收录战国文字资料时,我们对这两类形体进行了区分。如此处理主要基于两种考虑:一是两类形体差异很大,构形本义不同,前者是形声字,后者是会意字,而且都见于今天的楷书;二是这两类形体用法有别,各成体系⑫,《说文》中将后者当成前者的古文属于音近通假现象。将它们分列成两类字似乎更为合适。由于前后两次对数据的处理并不相同,而汇总数据时又未能予以统一,于是形成了两种标注,所以测试时标注成"御"字类形体,被模型识别成"驭",按照模型召回的结果,这种情况便被划分至识别错误的类别当中。除了这类学术方面的处理差异外,在标注过程中也存在偶尔误收、误标等情况,这些都会影响模型的识别。

(三)应用方面的尝试

为了更好地体验模型的识别功能,我们开发了应用平台,将智能模型部署在服务器上,试着转换成实际应用程序。同样,转化后应用程序的识别范围仍是覆盖6941个单字,识别的平均速度为38ms/张。举两个例子来观察实际应用情况。首先以  形为例,该形为"戒"字,见于戒鬲铭文(《铭图》2767)。将鬲铭拓本从输入端提交,会自动提示选择待识别的古文字形体。框选后页面右上角会有图片预览,确认无误后可点击确定,提交(参图4)。应用程序会很快给出识别结果(参图5),页面顶端显示的是识别字形,识别结果下面是"最优识别结果",为"戒"字,准确率达94.66%。考虑到古文字中存在很多形近字甚至是同形字,智能程序中输出结果除了"最优识别结果"外,还提供了参考结果项。"参考结果"是模型给出的其他可能性,共6个参考项,每个结果后面都以百分数标明了相似度并根据大小依次排列。例如参考结果中"武"字的相似度为"89.2%",排在最前面。除了两项楷书识别"结果"外,我们还设置了"相似字图"推荐功能。影响识别的因素中形近字占据很大比例,设置的"参考结果"可以在一定程度上解决形近字误认问题。一般情况下,测试两组形近字时,正确答案或者是"最优识别结果",或者是"参考结果"中的前两项,所以这一设置在一定程度上解决了形近字辨识的问题。

"戒"字见于今天的楷书,再举一个没有被保存下来的文字形体。如孟簋铭文中有形体作  (《铭图》05174),据研究此形从宀、从琮字初文得声[14],可隶定作"𡍮","𡍮"不见于今天的楷书,常用的字体库中未收录该形。如前文所论,我们曾对这类隶定字进行了录入与显示设置,所以应用程序可以正常输出该形。经过检测,应用程序最终也给出"𡍮"作为"最优识别结果",准确率达94.57%(参见图6)。

需要说明的是,应用程序中所设置的"最优识别结果""参考结果""相似字图"都可点击查看"详情",包括该字字形数据库、知识图谱、词义用法、古文字辞例等相关信息都可相互关联并能够直接查阅。这些功能涵盖

图4 上传图像

图5 选择字图

图6 识别结果

了文字的形体、用法、词义等诸多方面，能够为使用者提供很大便利。

三、古文字知识图谱的建设

目前的人工智能与古文字结合研究尚处于起步阶段。就古文字本身来说，研究的内容多集中于识别方面。这是技术在应用方面的实现，也是进一步研究的基础，那么当前人工智能技术能否为考证未释字提供一些帮助和提示呢?这也是我们在研究过程中一直思考的问题。以上文讨论的识别内容为例，其不但可以识别已释字，还可以为考证未释字提供参考。学者在考释古文字形体时，经常会将未释字与已释字加以系联，这种系联主要依赖于研究者的记忆积累和知识经验，而记忆联想与知识推理都是人工智能技术所擅长的。通过上述古文字识别任务，我们已将古文字字形图像分类至对应的文字标签，并在训练中使神经网络能够提取字图的关键特征。因此，我们可以通过提取神经网络输出的高维向量表示该字图，取所有字图向量的平均值，可得到表示字形的向量，通过计算向量之间欧氏距离以及余弦相似度等方法即可量化不同字形的相似程度。相似数值越接近，说明字形之间的联系越加紧密。这是从字形整体形态角度考虑，此外还可以计算字形所含偏旁及构形的相似度。这就需要引入人工智能里面的知识图谱技术。

(一)知识图谱与数据拆分

知识图谱本质上是一种语义网络，用于揭示事物之间的关系。在自然语言处理领域应用广泛，如语义搜索、智能问答、辅助决策等方面，其已经成为人工智能发展的重要动力。[15]知识图谱能够支持知识的抽取、融合、管理和应用等各个方面。我们将其应用在古文字研究中，一方面可以构建古文字知识网络；另一方面也可计算字形中偏旁及构形的相似度，进而寻找已释字和未释字之间的关联路径，为考证未释字提供帮助。

数据的收集和标注依然是研究的前提与基础，我们对数据集中古文字形体做了进一步处理。古文字中的偏旁部首是文字组成的基本单元，对古文字的考释有着重要作用。唐兰先生提出的偏旁分析法就是通过分析、识别偏旁来考释古文字，历史考证法也与此关系密切[16]，所以我们把数据集中的古文字形体进行了偏旁拆分，拆分时以古文字写法为基础，同一个字的不同写法予以区分处理，多数情况下会拆分到最小单位。如“御”

字作  ,可拆分成“亻、止、午、卩”四部分;另有异体作  ,可拆分成“止、午、卩、女”四部分。我们对数据集中古文字形体都做了分类和拆分并加以标注,形成了文字、字图、偏旁相互结合的数据集。

(二)知识图谱的构建与应用

利用已经完成偏旁拆分的古文字数据集并结合已有的字义数据集,我们设计并构建了古文字知识图谱[19],是首个关联甲骨文、金文、战国文字并对字形、偏旁、字义之间的关系进行描述和表示的历时知识谱系。图谱以单字为基础,将6941个单字与古文字字形、相应的偏旁相互关联,同时对应到当代楷书并系联到每个字的词义(架构图参见图7)。

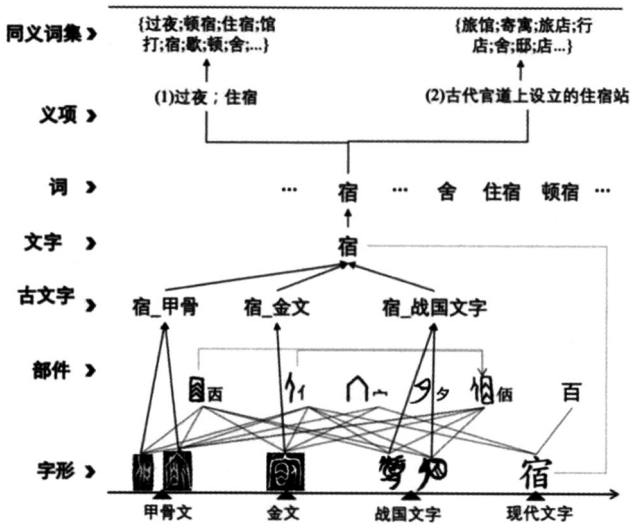

图7 古文字知识图谱架构图

古文字知识图谱目前分为七层,包括字形、部件(即偏旁)、古文字、文字、词汇、义项、同义词集。通过对实体进行属性描述以及层次内部与跨层次实体之间的关系描述,实现对古文字知识体系的多角度展示。其中重要的知识包括对应的古文字图像、偏旁、字义等相关信息,还可关联到释文辞例。此图谱是可以扩展的,字形层和古文字层对于每个特定时代的文字类型是相对独立的,随着数据的增加与补入,图谱信息也能相应地随之扩展。知识图谱可以非常直观地显示古文字知识,包括字形写法、部件组成、词汇字义等,也可通过部件等信息展现出不同字之间的关系。这些通过单字检索、偏旁检索便可实现。

古文字知识图谱还有另一项功用,可计算文字所含偏旁及构形方面的相似程度并进行量化。形体相近的古文字,在知识图谱中往往拥有更多的交叉,它们通过相同偏旁或含义相近的部件连接在一起(参见图8)。基于这种特性,我们采用随机游走算法,用向量来表示知识图谱中的字形实体,所得到的高维向量可以表示该字形在知识图谱中的语义空间。通过计算向量之间欧氏距离、余弦相似等方法即可量化不同字形在偏旁系统方面的相似程度。这种方式与古文字考释中的偏旁分析法有相似之处。

前文在利用计算机视觉算法实现文字的形体识别时,从字形的整体形态上可以计算出不同字形的相似程度,而这一部分讨论了知识图谱的构建,并且结合知识图谱可以计算出不同字形中偏旁及构形上的相似程度,把两种方法结合起来,通过计算得出的相似字形更加客观。通过以上方法,智能模型即可为我们推荐出与未

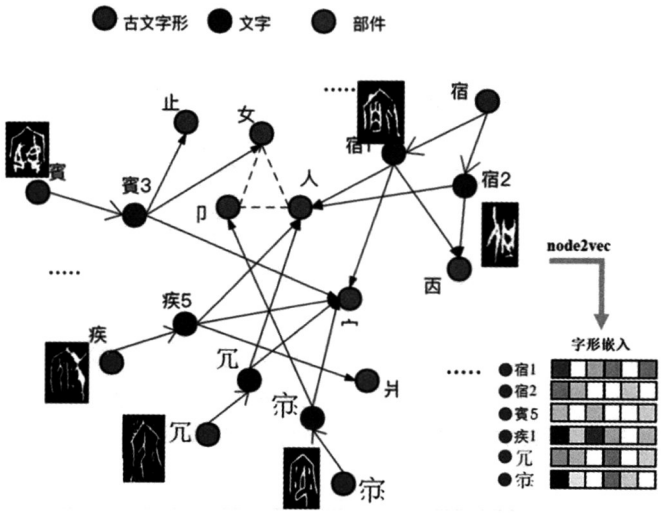

图8 知识图谱中形近古文字的关联案例图

释字相似的古文字形体，从而为古文字的考释提供一定的辅助。

本文讨论了人工智能与古文字相互结合的研究，利用深度学习算法来实现形体上的识别，同时也讨论了古文字知识图谱的构建。其实，无论是方法上还是角度上，古文字与人工智能的结合都有很大的研究空间。研究方法上，近些年自然语言信息处理技术发展迅速，在命名实体识别、语义关系等方面都有重大提升，可以利用文献数据训练智能模型，使其具有强大的古代汉语语感。在古文字考释过程中，输入待释字所在的文句，智能模型或者能够给出文献中的相似语句，或者根据前后文意、语词搭配缩小考释结果的范围。这些都能给研究者提供很大的帮助。研究角度上，除了古文字资料本身外，文字载体的研究也可结合人工智能技术开展。甲骨方面，学者利用人工智能技术缀合甲骨残片[16]，取得了不错的效果；青铜器方面，我们利用人工智能深度学习技术对青铜器进行了分期断代研究，实验结果和应用转化均十分理想。我们相信，随着更多学者的参与，人工智能与古文字的相互结合研究会取得更大进展。

注释：

①如部分传世金文和不够清晰的楚简材料。

②此数据集仍然在更新之中，目前正在整理的是竹简文字。

③中国社会科学院历史研究所：《甲骨文合集》，北京：中华书局，1978-1982年。本文简称《合集》。

④吴镇烽：《商周青铜器铭文暨图像集成》，上海：上海古籍出版社，2012年。本文简称《铭图》。

⑤罗福颐：《古玺汇编》，北京：文物出版社，1981年。本文简称《玺汇》。

⑥关于“膨胀”与“腐蚀”的概念，请参见杜慧敏、蒋忭忭、常立博等：《膨胀与腐蚀算法的改进及并行实现》，《西安邮电大学学报》，2017年1期。

⑦其实“为”字与  形存在一定区别，此次误释应属偶然。

⑧这一工作我们也做了尝试，不同材料呈现出的结果存在差异。

⑨或怀疑此铭是在封地所作，作器者来自距离周王畿较远的诸侯国。参见杨安：《释伯鼎铭文中的“茵”》，《古文字研究》第33

辑，北京：中华书局，2020年。

⑩如甲骨文中“鬼”字多从跪坐人形，也偶有写成从女者。陈剑先生认为它们不是表意相近的偏旁替换现象，应该理解成跪坐人形上增加了“敛手”的特征而与“女”写法相同。参见陈剑：《卜辞｛凶｝词觅踪》，《首届出土文献语言文字研究国际学术研讨会》，第18页注释2，台北：彰化师范大学等，2022年。

⑪另一例见于姒丁尊(《铭图》11797)。

⑫两者用法上的区别曾有多位学者提及，相关工具书亦论证颇详，参见张世超、孙凌安、金国泰等：《金文形义通解》，京都：中文出版社，1996年，第391-399页。

⑬历史考证法，是根据不同时期形体特征及演变规律来考释古文字的，这些形体特征和演变规律很多体现在偏旁部件的写法上。

⑭关于图谱的构建过程，请参见迟杨、Fausto Giunchiglia、史大千等：ZiNet：Linking Chinese characters spanning three thousand years(跨越三千年的汉字知识图谱构建)，Dublin(都柏林)：ACL(The Association for Computational Linguistics，国际计算语言学协会)，2022年。在这篇文章中，此知识图谱被称作“ZiNet”。

参考文献：

[1]唐兰：《古文字学导论》，济南：齐鲁书社，1981年.

[2]于省吾：《于省吾著作集·甲骨文字释林·序》，北京：中华书局，2009年.

[3]林沄：《古文字研究简论》，长春：吉林大学出版社，1986年.

[4]何琳仪：《战国文字通论(订补)》，南京：江苏教育出版社，2003年.

[5]滕壬生：《楚系简帛文字编》，增订本，武汉：湖北教育出版社，2008年.

[6]马承源主编：《上海博物馆藏战国楚竹书(二)》，上海：上海古籍出版社，2002年.

[7]荆门市博物馆：《郭店楚墓竹简》，北京：文物出版社，1998年.

[8]马承源主编：《上海博物馆藏战国楚竹书(一)》，上海：上海古籍出版社，2001年.

[9]湖北省荆沙铁路考古队：《包山楚简》，北京：文物出版社，1991年.

[10]蒋玉斌：《释甲骨文“烈风”——兼说“𢦏”形来源》，《出土文献与古文字研究》第6辑，上海：上海古籍出版社，2015年.

[11]许慎撰、徐铉校订：《说文解字》，北京：中华书局，2013年.

[12]容庚著，张振林、马国权摹补：《金文编》，北京：中华书局，1985年.

[13]董莲池编著：《新金文编》，北京：作家出版社，2011年.

[14]陈剑：《释“琮”及相关诸字》，《甲骨金文考释论集》，北京：线装书局，2007年.

[15]张吉祥、张详森等：《知识图谱构建技术综述》，《计算机工程》，2022年3期.

[16]莫伯峰、张重生、门艺：《AI缀合中的人机耦合》，《出土文献》，2021年1期.

