系列笔谈之四：智能时代古籍OCR技术

北京大学数字人文研究中心

王军（北京大学信息管理系）：OCR 技术是古籍数字化技术的核心和基础

OCR技术是古籍数字化技术的核心和基础，现代汉语常用汉字约七八千字，而古籍中所包含的文字最高达8万。由于文字量庞大，异体字众多，字形多变，版式多样，而且年代久远，页面模糊，再加上缺乏充足的训练数据，这使得古籍OCR比一般的OCR任务更具有挑战性。近年来，深度学习技术在这个领域的应用显著提高了OCR的准确率，大大降低了应用门槛。它现在是古籍数字化领域受关注度最高，应用面最广，需求量最大的一门技术。

我们非常荣幸地邀请到了古籍OCR领域的顶级专家——中国科学院自动化研究所副所长、模式识别国家重点实验室主任、中国科学院大学人工智能学院副院长刘成林教授，中国图象图形学学会常务理事、华南理工大学电子与信息学院教授、文档图像分析与识别专委会主任金连文老师，以及安阳师范学院计算机与信息工程学院院长、甲骨文信息处理教育部重点实验室主任刘永革老师。三位老师向大家全面介绍古籍OCR的基础知识、技术流程和前沿进展。此外，我们还特别邀请了在古籍OCR领域的业界翘楚书同文公司总裁张弛宜女士，向大家介绍他们在古籍OCR领域二十多年的业界实践。

刘成林（中国科学院自动化研究所）：复杂文档图像版面分析

古籍OCR现在受到学术界和业界的高度关注，但是相对于人工智能领域其他方向的进展，尤其相对一般的OCR而言，古籍OCR的技术还是相对滞后的。其中一个因素是，学术界研究一般的OCR和一般的模式识别技术比较多，但专门针对古籍的研究相对较少，而古籍数字化业务工作者与人工智能学术界的交往也不多。

古籍文档有一些专门的特点，存在其特定的难度，包括版面复杂、字符类别数较多、文档图像质量差等问题。其中，版面分析是非常关键的一个问题。文档的版式变化特别多，古籍文档的版面又有一些特殊性，比如古籍的文本行之间挨得非常近，而且存在大字行和小字行相互交织在一起的情况，很难将其分清。我主要研究一般复杂文档的版面分析，今天我就分享一下关于复杂文档图像版面分析技术的最近研究进展，希望这方面的技术对古籍OCR能有些借鉴。我相信，古籍跟一般文档版面分析的技术应该是相通的。

文档图像分析和识别的流程主要有四个阶段：通过扫描或者拍照获得文档图像后，第一个阶段是图像预处理，包括去噪、形变矫正、去背景等。古籍文档图像的噪声比较多，去背景是比较困难的。第二个阶段是版面分析，主要是把图像按照不同内容划分成不同区域，比如以是否有印章、文本行、图形、公式、表格等为标准划分区域。第三个阶段是文字识别，现在一般是以文本行为单位（古籍一般为纵向排列，称为文本列，我们统称为文本行），对文本行里的字符同时进行切分和识别。有些方法只给出识别结果，不提供切分结果，但切分结果对于后续的置信度估计和后处理是非常关键的。第四个阶段就是后处理，对文本识别结果进行检查校对，或基于识别结果提取语义信息。

我顺便也简单介绍一下文档分析与识别（也就是OCR）的研究历史，其实这个领域的研究历史与模式识别及人工智能的历史差不多。自1950年代计算机科学开始发展之后，人工智能和模式识别逐渐形成了一套理论和方法体系。在文字识别领域，即OCR领域，它的方法最早主要是统计模式识别，后来到句法模式识别、人工神经网络、支持向量机等，发展到现在的深度学习。OCR识别的对象早期主要是单字识别，后来慢慢扩展到英文的单词识别、中文的文本行识别，到最近在探索的整页识别。字体或书写风格也从早期的单字体到多字体，印刷体到手写体，等等。相对来说，古籍文档大概是2000年以后才开始受到较多关注。现在，古籍OCR技术进步比较明显，但是相对一般的OCR而言，古籍的样本数

比较少，即使深度学习技术也难以达到非常高的识别精度，所以其实际应用仍然需要较多的人工后处理。

下面进入我的正题——文档图像版面分析。古籍的版面是非常复杂的，文本行或文本列并不整齐，挨得也比较近。其次，古籍中一些大字的文本行和小字的文本行挨得较近，很难分开。版面分析从1970年代开始已有较多研究。早期的方法多基于一些规则，比如投影分析等，这类方法被称为自上而下的方法——将文档从大块逐步切成小块，如果它是比较规则的，比如矩形块，那么自上而下的方法就可以分得较好。另外一类是针对不规则文档，采用自下而上的方法——基于聚类的方法，比如对像素或者对连通部件，经过聚类逐步得到文本行再得到文本区域等。大概2000年之后，发展了很多比较先进的方法，以针对更加复杂或者变形的文档。这一类方法包括变形模型的方法，比如主动轮廓模型、基于图的聚类方法等。这种聚类可以更加灵活，可以得到不规则的文本区域等。

现在基于深度学习的版面分析方法大概有三类。第一类是类似于一般的物体检测的方法，如Faster R-CNN，对文档区域或文本行直接进行检测。针对非规则文本行也有相应的方法，如场景文本检测中对于任意方向文本行、任意形状文本行的检测。第二类方法更加灵活，是基于实例分割的方法。实例分割基于像素分类或像素聚类，所以非常灵活，其代表性方法是全卷积神经网络，较适合做像素级分割。第三类方法是深度学习结合关系推理的方法，如像素分类、连通部件分类和聚类等，可以用图神经网络进行分类或聚类，同时考虑像素或连通部件之间的关系。

我重点要讲的是最近几年非常有效的实例分割和关系推理两类方法。关于版面分析，有一些较常用的文档图像数据库。现在公开的古籍文档数据库还比较少，欧洲发布了一些数据库，他们称为历史文档，其实那些文档并不是特别古老，最多是几百年前的。到现在为止，还没有见到特别大规模的中文古籍文档公开数据集，今后需要加强这方面的工作。现在公开的早期文档图像数据集，大概是2000年前后美国华盛顿大学发表的扫描印刷文档数据集；2000年之后，手写文档数据集陆续公布，包括手写文本行分割的数据集。欧洲学者发布了几个比较复杂的文档数据集：一是2014年法国A2iA公司（专门做OCR的公司）发布的包括法文、英文和阿拉伯文的多语言文档图像数据库Maurdor。其文档数量特别大，总计几万张图片，版面非常复杂，且很多文档是手写体和印刷体混合的。二是在2017年“国际文档分析与识别会议”（ICDAR）开展的手写文本行分割的竞赛中公布的cBAD数据集。它主要面向文本行的分割，重点在检测出每个文本行的基线。三是在2017年瑞士弗莱堡大学举行的竞赛中发布的中世纪文档的数据

集。其文档给出了像素级的标注，即每个像素标注类别，如属于背景、文本或注释等。因为像素级标注的工作量非常大，所以文档数量不多，三类文档每一类只有40页，20页用于训练，20页用于测试。

我们研究组提了两个版面分析方法。一是基于实例分割的方法。计算机视觉领域对实例分割的方法研究很多，发展很快。比如，要将像素分成前景和背景两类，实际上是一个二值化问题。如果有多类物体，就要对每个像素做多类分类。典型的方法是用全卷积神经网络（FCN），其输入层和输出层都是相同大小的图像，输出层就相当于给出了每个像素的分类结果。其中一个目前使用较多的代表性方法——把整个编码层和解码层之间进行连接以此来提高特征的使用效率，被称为U-Net。全卷积神经网络用于文档分割，我看到一篇较早的文献是2016年华中科技大学团队基于像素分类方法将FCN用于场景文本检测。[①]为了将场景文本行检测出来，先将文本行框里面的像素标记为前景，其余为背景。然后，基于这样一些标记训练神经网络，就可以把前景区域的像素检测出来。后面还有关于字符的检测，即在训练图像里把每个字符的中心点区域标记出来，训练神经网络就可以检测字符的中心位置，基于中心位置，得到比较准确的文本行。几乎同期，大概在2016—2018年，一系列的工作将全卷积神经网络用于手写文档的文本行分割。2017年ICDAR文本行分割的竞赛上也有些方法使用了全卷积神经网络。

G. Renton等人在2018年IJDAR（《国际文档分析与识别期刊》）发表的文章中，其分割方法就是用FCN对文本行的中心区域进行预测。[②]训练图像过程中一般需要标注文本行的中心区域：中心区域像素标为1（前景），其余像素标为0（背景），这样就可以用全卷积网络将文本行的中心区预测出来，然后再进行一些后处理，把每个文本行的像素和它的中心区域相连，得到完整的文本行。这个卷积神经网络用到了一些所谓的膨胀的卷积核，因为一般的卷积神经网络在每个卷积层之后要做降采样。利用膨胀的卷积核，卷积核可以越来越大，不需要降采样也能保证它的感受野是越来越大的。当时实验结果显示膨胀卷积核神经网络比一般的卷积神经网络的分割性能更好。最好的性能是在2017年竞赛上欧洲几个大学发布的结果。这个结果也使用了全卷积神经网络U-Net，但是像素分类后的后处理做得比较细致，所以最后性能较高。这是文本行分割的精度、召回率和F值。

---

①Zheng Zhang et al., "Multi-Oriented Text Detection with Fully Convolutional Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), LasVegas, June 2016, DOI:10.1109/CVPR.2016.451.

②G. Renton et al., "Fully Convolutional Network with Dilated Convolutions for Handwritten Text Line Sementation," *International Journal on Document Analysis and Recognition(IJDAR)*, vol. 21, no. 3, 2018, pp. 177-186.

这篇文章里U-net得到的F值是91%，不是最好的，说明没有把太多功夫花在后处理上。这个工作也就说明用全卷积神经网络基于实例分割的方法做手写文档文本行分割是比较有效的方法，当然也可以用于其他版面区域的分割。

我们实验室发表在2018年国际人工智能联合大会（IJCAI）上的多任务学习的全卷积神经网络方法，也利用U-Net来做版面分析。多任务是因为文档中有多类要素区域，也就有多种版面分割问题。比如古籍文档，背景比较复杂，有很多噪声，它的二值化是一个较难的分割问题，这也用全卷积神经网络来解决。其次，要将文字和注释区域区分开，这又是个分类的问题。同时，文本行分割也是一个分类问题，还有文本行的检测，文本行中心区域或基元的检测等。我们设计的神经网络上有四个输出任务模块来进行多任务学习。实验数据是瑞士弗莱堡大学发布的中世纪文档数据集，只有三类文档，每类文档20张图片用于训练，20张用于测试。分割的不同区域基于像素级进行评价，即每一类像素分类的召回率、精度和F值。和2017年ICDAR竞赛的最好结果相比，我们的方法性能提升比较明显。文本行检测的性能用文本行级别的IoU重叠度进行评价，文本行的基线检测结果说明全卷积神经网络可以用于多任务学习的版面分析。举例来说，原图像经过多任务的神经网络，得到文本行分割结果（把不同区域的文字标记成不同的颜色）和基线检测的结果。

关于分本行实例分割，我们又提出另一个方法。因为有些文档不同文本行之间挨得非常近，用一般的全卷积神经网络很难分开，所以我们提出基于标记金字塔网络的实例分割方法。[①]这个标记金字塔是什么呢？如果对每个文本行给出四层标记：将文本行核心区域标记为4；核心区域稍微扩大一点，标记为3；再扩大一点，标记为2；整个文本行区域标记为1。如此将四个层次的标记，当作四个任务进行训练。训练之后，神经网络能对这个文本行的分割给出不同层次的标记，把四个层次标记得到的map合并之后，那么就会得到一个置信度图，这个文本的核心区域与外围区域的置信度是不一样的。在这个图上，再用一般的图像分割方法，比如分水岭算法，就可以把文本行很好地分开。根据一些实验结果的例子，针对文本行之间挨得很近的情况，用一般的全卷积网络没法分开，用标记金字塔网络就可以分开得很好。基于中国科学院自动化研究所的手写文档数据集做的比较实验中，如果人为地把不同的文本行拉近，比如拉近10个像素，甚至50个像素，那么一般的实例分割方法的性能下降得非常快，标记金字塔网络即

①Xiaohui Li et al., “Instance Aware Document Image Segmentation using Label Pyramid Networks and Deep Watershed Transformation,” International Conference on Document Analysis and Recognition (ICDAR), Sydney, September 2019, DOI:10.1109/ICDAR.2019.00088.

使在拉近50个像素的情况下，其文本行分割的正确率还可以达到80%以上。至于英文数据集，我们在2017年ICDAR竞赛的文本行数据集上做了文本行基线检测。基于Maurdor多语言文档数据集做了区域分割的实验，其中有些指标类似于聚类，与过去相比，性能提升也是非常明显的。

以上几个基于实例分割的版面分析方法是非常有效的，但是实例分割方法需要像素级的标注。像素级标注有一些技巧，比如我们把文本行或者区域，用矩形或者多边形轮廓标记出来后，可以把这个区域内的像素都标为正类，其余的标为负类。基于这样的像素分类，大部分情况下可以分割得比较好。但有时如果没有充分利用不同区域之间的关系，分割的精度会不太高。所以，最近几年基于图神经网络的关系推理方法发展得很快。图神经网络是基于图结构，可以表示文档里面不同区域之间的关系。比如这个文档里面有很多连通部件，需要判断哪些部件是文本，哪些是背景，而且哪些连通部件应该连成一行。这就可以用一个图来对这些连通部件同时进行分类，并判断它们之间的连接关系，比如相邻连通部件是否属于同一行。过去传统的方法是用条件随机场，现在一般用图神经网络。图神经网络先是用在联机手写文档上面，因为联机手写文档可以以笔画为单位，不像文档图像必须要先通过一些预处理提取连通部件。联机手写文档以笔画为单位进行分类或聚类，以此将文本行和不同图形区域分开。

图神经网络是如何用于版面分析的？图神经网络是一个图（graph），可以通过节点和边分类，来对节点进行聚合。而且，图神经网络是一个多层网络，它的节点和边的特征经过传播之后，可以把局部的节点和边的信息，即上下文信息，进行整合，而且它的参数可以自学习。一个比较有代表性的方法是图卷积网络和图注意网络。因为图的拓扑结构变化多，图注意网络的一个特点就是相邻两层之间传播的权值在所有的节点之间是可以共享的，那么它就跟拓扑结构无关。真正传播权值是根据注意力的权重来决定的。比如当前这个节点跟前一层邻域节点之间计算相关度，根据注意力机制把相关度做softmax归一化得到注意力权重，再根据注意力权重来计算两层之间的特征传播。每一层可以有多头注意力，相当于是多个特征的map，那么最后在输出层的每个节点根据特征向量直接进行分类就行了。如果有边的特征的话，可以对边进行分类。比如这两个节点是否应该连起来，是一个两类的边分类问题。

为了更好地做版面分析，我们提出了边注意网络。因为常规的图注意网络只有节点特征的传播，没有边特征的传播。为了做边特征传播，我们使用边注意网络，把边特征加进去重新计算注意力的权重，由此，节点特征和边特征可以同时进行传播，这样可以更好地利用图像里面的上下文信息。

基于边注意网络，我们在联机手写文档方面做了一系列工作。其一就是联机手写文档的笔画类型的分类。在文档里面，可以将流程图区域、图形区域、表格、文字分成不同类别，相当于将图像分成不同区域。我们提出这个方法可以对文本行进行进一步的分割，叫边聚合的注意网络，还可以用于流程图的识别。

脱机手写文档怎么做？刚才讲了脱机手写文档要基于连通部件的聚类，可以使用图神经网络。图神经网络输入的底层特征，就是用全卷积神经网络提取特征之后，提取一些连通部件构图。构图之后，把从每个节点和每条边提取到的卷积特征，用图神经网络进行多层传播，最后用条件随机场进行分类。我们也注意到添加条件随机场后，性能变化不是很明显，但条件随机场还是能进一步提升性能。对于印刷文档而言，连通部件的提取比较方便。基于文档图像二值化，我们做了水平方向连通部件聚合，减少部件的数量。每个连通部件从卷积特征平面把对应区域内的特征聚合起来得到节点特征。如果是两个连通部件，就把它们合并区域里的特征提取出来作为边的特征。在图神经网络最后一层对节点和边分类，在此基础上把连通部件聚合成文档区域。我们在一个公开的大规模印刷文档数据集PubLayNet上进行实验，最终区域分割的精度超过97%。

同时，我们也把基于图神经网络的方法应用于文档的表格检测与结构识别。有表格的文档版面是比较复杂的，尤其是表格的结构分析，需要把表格里面每个表格单元检测出来且判断相邻单元之间的关系：是否属于同一行同一列，是否标记项或填充项等。首先，我们用基于实例分割的方法使用全卷积网络提取特征，检测连通部件。以连通部件为节点构图用图神经网络检测表格或者检测表格单元。然后以表格单元为节点再构成一个图，形成第二个图神经网络，用来做表格单元之间的关系推理。在2019年ICDAR北京大学组织的表格检测与识别竞赛公布的数据集上，当时最好的结果是表格单元关系分析（表格结构识别），但其正确率仅达到0.2%以上，这是非常低的，而我们的方法把它提升到0.7%以上。

下面介绍一下我们在古籍文档方面做的一项工作。2018年，我们构建了一个大规模的文档数据集。当时古籍文档数据集比较缺乏，我们买了几十本书，包括《四库全书》中的25册、古代十个时期的经文389卷，然后对这些材料进行扫描。《四库全书》写得比较工整，总共7,000多页，经文4,000多页，共计1万多页。将文档扫描之后，我们把这个文档所有的文本行、不同的区域，以及每个字符的位置、类别都标记了出来。这个数据集的字符样本集在2019年已经发布了，文档页面的数据集还没有正式发布，但未来仍需要把它充分利用起来。为了标记这些数据集，我们用多任务的全卷积神经网络，对文档做二值化区域分割、文本行分割和字符切分。字符切分，就是对字符中心进行预测，然后将字符位置

与标记文本对齐，就可以得到每个字符图像的类别标号。从对《四库全书》和经文两种风格的文档分别标注了的100页训练图像和100页测试图像来看，我们进行的页面分割、文本分割、字符切分的精度还是比较高的，少量的错误由人工进行校对修正。

我们将所有的切分字符图像单独拿出来，制作了数据集并发布。数据分两种风格——《四库全书》(风格一)和经文(风格二)。所有的字符类别数加起来有12,000多类，但是大部分字符集中在这两种风格的文档里，共有的类别为2,365类。还有一些出现频率较高的类别，我们把它作为风格一与风格二的加强集。另外，还有几千类的字符数据样本数特别少，每一类不到十个样本，我们还没有公开展示。

基于这个古籍字符数据集，我们做了一些初步的实验，两种风格分别用卷积神经网络或卷积原型网络进行分类。风格一用卷积神经网络分类，精度可以达到98%以上；如果把基于风格一训练的神经网络直接用于风格二的样本分类，精度就很低。这说明这两种文档的风格差异是较大的。这样一个数据集，在人工智能研究领域，可以做迁移学习、类别增量学习等研究。在将来的古籍OCR领域，这样的技术也是非常重要的。

现在文档分析技术整体上进步非常大，手写文档识别现在已达到实用程度，但是古籍文档还是相对滞后一些，因为它缺乏大型的数据集，尤其像现在深度学习的方法都需要大量数据来训练。未来的重要工作首先就是要构建大规模的图像和文本语料库，并组织一些竞赛来推动这个领域的发展。我们也可以参照人工智能领域构建针对古籍文档的预训练大模型，包括跨模态的大模型。在技术研究方面，图像预处理、版面分析、文本行识别、置信度估计等都还有很多需要研究的问题。机器学习里面的很多方法，包括小样本学习、迁移学习、持续学习、交互学习、跨模态学习等，在古籍OCR领域有重要作用。因为现在样本数不太多，并且样本是不断增加的，识别之后还要进行交互式后处理等，所以这些学习方法的研究非常有意义。从应用的角度来说，古籍文档的语义提取、信息检索、关系推理、文档大数据分析，将来会有很大的应用前景，可能会在不同领域推广应用，如图书馆、档案馆、出版社，及家谱、地理志的编修等。其实绝大部分的古籍文档还处于纸张或图像状态，还没有被识别，没有被充分利用。

金连文（华南理工大学电子与信息学院）：古籍 OCR 的数据、方法和应用

中国有几千年的辉煌文明，在上下五千年的历史长河中，传承了非常多的珍贵古籍。这些古籍承载着丰富的历史、文化、政治、经济等方面的信息，具有非常大的价值。据2016年国家图书馆牵头启动的“全国古籍普查登记基本数据库”项目数据，截至2020年11月，普查统计的中国古籍图书达790多万册，虽然可能有重叠部分，但是从这一数量也可以看到，我国从古到今流传的古籍非常多。在古籍数字化进程中，四十多年来，我国已完成很多大型的古籍数字化项目，如“《四库全书》(文渊阁电子版)”“中国基本古籍库”等代表性工程。但到目前为止，依然有大量的、各种各样的古籍，还没有完成文字识别转录等工作，这是由于有不少古籍的版面非常复杂，现在的技术虽然已经可以很好地解决简单板式的文字识别，但是复杂版式的古籍，如家谱、地方志等，要将其自动化地、高精度地识别出来，还是很困难的。如果只是以影印的方式进行古籍数字化，不便于阅读、编辑、检索。因此，利用光学字符识别技术（简称OCR），可以帮助我们更好地识别古籍中的内容、文字，分析版面并进行结构化输出，这对于古籍保护、检索，乃至信息挖掘和知识发现，都有非常重大的意义。

但古籍文字识别仍然是一个极具挑战的问题，主要表现在四个方面：第一，古籍版式很复杂，文字密集、图文混排，给分析带来较大困难；第二，不同朝代的刻字工匠书写风格差异很大；第三，有些古籍存在着严重的图像质量退化问题，如残缺、模糊、背景噪声、污渍干扰等；第四，古籍文字的类别数巨大，《康熙字典》包含近五万字，目前几乎还没有这样一个完整的标注的古籍图像数据集，要识别这样大规格的数据集也很困难。另外，古籍中有大量异体字，将异体字很好地进行区分也并不容易。

与常规OCR技术相比，古籍OCR的困难和挑战主要表现在古籍的质量、版式、风格等方面。目前主流的OCR技术对于印刷体文字的识别率已经很高了，但它不能直接应用于古籍OCR，主要原因在于目前缺乏高质量、大规模标注的数据，特别是中文的大规模公开数据集很少，版式数据集大部分也是西方古籍的。

日前，我们挑选了《敦煌遗书》中的一些古籍，利用国内几家较知名的大厂的OCR引擎进行了测试，发现识别结果很差，文字检测经常出错，甚至看起来不是特别难的问题也会出错。这倒不一定说明它们的技术不好，我认为通用的OCR引擎在没有对古籍数据进行针对性训练的情况下，是不能够达到较高水

平的。国内也有做得比较好的平台，如书同文公司的“i-慧眼OCR”效果就还不错，当然偶尔也有少量错误。我们团队也开发了一个大藏经的古籍OCR识别引擎，这个引擎主要以大藏经数据来训练，对《敦煌遗书》的识别率还不错。

今天如果我们有质量好的标注数据，利用人工智能技术，特别是以深度学习为代表的人工智能技术，可以比较好地解决古籍OCR问题。事实上，在文字识别领域，以神经网络为代表的人工智能技术已有多年研究历史。比如在深度学习人工智能时代，应用最多的卷积神经网络、循环神经网络模型，在早期都大量应用于文字识别中。1998年深度学习领域的国际权威专家Yann LeCun在贝尔实验室研发的手写体数字识别系统，第一次较正式地提出了卷积神经网络的模型——LeNet5，这个模型在当时已达到很好的识别效果。此工作的影响是非常深远的，到目前为止引用超4万余次。2020年麻省理工学院的《科技创业周刊》（*MIT Technology Review*）采访了人工智能领域国际权威学者Hinton教授，他当时提出一个观点：深度学习将无处不在、无所不能。这个观点虽然有些激进，但在某种意义上可侧面反映出以深度学习为代表的人工智能技术，确实对今天很多学科领域，包括计算机视觉、多媒体数据、自然语言处理，乃至教育、金融、医学、生物、电子商务、考古等，都发挥着重要的作用。

为什么深度学习会发挥如此重要的作用呢？第一，现在有充足的大数据，尤其是在互联网时代，数据呈爆炸式的指数增长。第二，我们有很好的计算能力，如以GPU为代表的大规模并行计算硬件飞速发展。第三，在算法层面有很多突破，包括正则化防止过拟合学习的方法与技术，在识别、检测、分割、序列建模，乃至大规模预训练、自监督学习方面，都有非常多突破性的、有影响力的工作，每年都会有很多令人眼前一亮的进展。所以我相信，利用人工智能的一些新技术，可以把古籍OCR问题——从古籍图像处理、版面分析、文字检测到文字识别，到文字的结构化输出，很好地解决。

大家认为人工智能主要有三大要素：数据、算法和算力。当然也有学者表示，除此之外还有一个因素，即知识。所谓知识，包括领域知识、语义知识、物理常识、世界知识、无监督预训练知识等，可以帮助解决各种人工智能问题。如果有一个大规模预训练好的模型，有很好的语言语义方面的知识，可以很好地帮助我们解决古籍OCR问题。受时间和主要关注点所限，我主要从传统的三个方面进行介绍。

（一）目前国内外公开的一些古籍数据集

数据集是人工智能的一个重要因素，如果有海量的、高质量的标注数据，

古籍OCR并不是太大的难题。但我们在2016年下半年做古籍OCR相关调研时发现，国内公开的古籍数据集比较少，尤其是中文的数据集，所以我们团队在2017年左右，构建了一个高丽藏数据集和多样式的大藏经数据集，分别是TKH和MTH。同刘成林教授团队2019年提出的数据集相比，我们的单字规模要小一些，但这两个数据集是篇幅级古籍文档数据集，有文本行、篇幅、版面等标注数据，总计超过100万字，覆盖的汉字类别也超过了几千类。从我们最新构建的MTH第二个版本的数据集（MTH v2）中的一些样式可以看到，它覆盖不同的版式、字体、排版风格等，这个数据集含有不少双列夹注等识别解析的难点。另外，我们进行了字符集级别的标注，同时提供文本行以及文字的阅读顺序的标注，所以信息是非常丰富的。第二个数据集是中科院自动化研究所的“CASIA-AHCBD数据集”。这个数据集规模非常大，类别数也非常多，代表性较强，涵盖了“四书五经”和佛经，是国内目前已公开的较大规模的中文古籍单字数据集。第三个数据集“HDRC-Chinese”，2019年ICDAR OCR竞赛中由新加坡的一家公益机构提供，是以中文家谱为重点的数据集。此数据集在ICDAR 2019用于三个比赛，包括家谱的文本行识别、版面分析及文本检测识别。家谱版式复杂，有些排版并不规范，所以难点在于怎样更好地做版面分析。第四是甲骨文的数据集，其中三个代表性的是Oracle-20K、Oracle AYNU和OBC306。

外文数据集比较多。较早的版式数据集有：1.IAM-HistDB，由三个子集构成的德文版古籍数据集。2.DIVA-HisDB，中世纪的手稿数据集。3.HJDataset，日本的一个复杂版面分析数据集（日文古籍中的很多文字与汉字相似。此数据集的规模尚可，具备一些版式分析元素）。4.READ-BAD，欧洲历史档案的手稿数据集，大约包含2,000多张图片。5.REID2019，印度的一个小规模数据集，仅包含几十张图片。6.MapSeg2021，古代地图的历史文档数据集。古代的地图具有考古和学术价值，理解地图中的文字很重要，但其文字排布大多杂乱无章，将其很好地理解和识别出来是很困难的。MapSeg2021是一个巴黎地图数据集，在2021年OCR领域主流国际会议ICDAR中公开。此外，我们团队也正在建设一个包含复杂版式的古籍版式数据集，将古籍版式定义为27个类别，包括书眉、版框、卷、标题、正文、夹注、插图等。这个数据集的特点是：数据多样性，囊括《四库全书》、国家图书馆的古籍善本、佛经等；装潢多样性，收录卷轴装、包背装、影印本等不同版式；字体多样性，涵盖手写体、印刷体等不同字体。这个数据集已正式发表在ICFHR 2022，[①]可公开下载。

①Xiaoyi Zhen, Cheng Jiang, Shang Wu, Lianwen Jin, "SCUT-CAB: A New Benchmark Dataset of Ancient Chinese Books with Complex Layouts for Document Layout Analysis," ICFHR 2022, pp. 436-451.

（二）古籍OCR相关的基础性技术

OCR的一般流程主要包括版面分析、文本行检测、分割字符、文本行识别、字符识别及后处理。这是光学文字识别的一般流程，古籍OCR也大致遵循这样的流程。近年来OCR特别是场景文字识别是研究热点问题之一，虽然通用OCR或场景文字识别的研究很多，但是古籍方面的研究及报道并不多。因此，我从五个方面介绍古籍OCR涉及的一些技术，主要包括：（1）文字检测所用到的一些背景知识、物体检测和场景文本检测方法；（2）高精度的古籍文字分割和检测方法；（3）古籍文字的识别方法；（4）古籍的版面分析和端到端的识别；（5）古籍文档的阅读程序和理解。

首先是背景知识。在文字检测中用到的计算机视觉当中的目标检测方法，代表性的目标检测方法有一阶段、两阶段两大类方法。典型的一阶段方法有YOLO、SSD等，两阶段方法有Faster RCNN、Fast RCNN等。

一阶段的目标检测的方法中，普遍认为YOLO是一个又快又好的方法，它是在2016年CVPR提出的一个框架，用全卷积的网络直接去回归待检测物体的包围框坐标，最后通过后处理得到检测结果。YOLO-v2版本做了一定改进，主要体现在正则化技术的运用，引入Anchor机制，提升了检测精度。YOLO-v3版本搭建了一个更好的主干网络，用聚类方法可得到更好的Anchor尺度的设置等。YOLO系列仍在持续改进，如YOLO-v4/v5是最近开源出的一个模型，引入了更强的主干网、Bag of Freebies(BoF)、Bag of Specials(BoS)、CIoU等技术来提升检测精度。

两阶段的目标检测方法，比较典型的是Faster R-CNN，它用CNN特征提取了Region Proposal Network（简称RPN网络）得到一系列候选框，再通过CNN两个头的分支，一个做物体的bounding boxes回归，一个做物体分类，最后得到目标检测结果。Mask R-CNN在Faster R-CNN的基础上做了进一步改进。其一是用RoI Align替换RoI Pooling，可使进行下采样时，减少采样误差。其二是增加了一个实例分割分支，名为Mask分支。虽然改进非常小，但它可以更加细腻地进行物体的分割和检测，是一个简洁且十分有效的方法，获得了ICCV 2017的最佳论文奖。

在文字检测领域，也有非常多的特定的文字检测方法，近年来的研究热点主要是针对复杂场景文本检测（特别是弯曲文本）提取出来的。古籍中的弯曲文本相对少见，下面我介绍几个针对矩形框的文字的代表性方法。一个早期的代表性方法是在2017年CVPR提出的EAST模型，这个方法采用PVA主干网络，预测要检测文字的框、每个点到矩形框顶点的位置和角度值，可以较好地实现多方向的文字检测。同年另一个比较简单实用的方法——TextBoxes，它基于SSD目标检

测框架提出来的文字检测方法。TextBoxes针对文字检测问题，对Anchor做了一些特殊设计，效果也还可以，是早期开源的文本检测方法之一。

我们团队在2017年也提了一个多方向的四边形文本检测框DMPNet，这个方法同样基于SSD，但应用了蒙特卡罗方法匹配计算，并提出一些新的优化函数，实现了精准的矩形框回归。基于此方法，我们在ICDAR 2017年的多语言文本检测比赛（MLT）中拿到了第一名。此后，我们在2019年还提了一个多方向文本检测框的检测方法BDN，这个方法是基于Mask R-CNN改进的，只增加了一个关键边的设计分支，较好地解决了在数据标注过程中存在顺序标注的歧义性问题，在2019年ICDAR的街景文本检测竞赛中取得了第一名。这个方法也开源出去了，对于密集型的文字检测效果还不错。

第二是古籍文字分割和检测的一些方法。能够精准地把古代的文字从古籍中切分出来，是很有意义的。首先，它可以帮助我们更好地保护文物、文字。其次，将来把OCR数据建好，识别出的结果变成可编辑的文字后，有时还需回溯到最原始的第一手资料所存在的位置，高精准的文字检测，可以帮助我们进行文字溯源；再次，高精准文字检测也更有利于后续的文字识别。此外，把文字切分出来，可以方便文字学专家研究古文字的演变。所谓高精准，即检测框跟机器真实的标注框的重合度要足够大。在计算机视觉、人工智能领域中，一般认为IoU[①]=0.5，就算检测正确了。但我们注意到不少古籍文字如果按IoU=0.5来切分，字是不完整的，只有到0.7以上甚至到0.8，才能把字较完整地保留下来，所以高精准的文字检测就变得非常有意义。

大约在2017—2018年，我们针对文本检测和分割的问题，基于传统方法做了几种的尝试。因为当时标注数据非常少，所以我们设计的方法是在没有文本框的标注情况下，先利用直方图投影获得框的大致位置，再做标注框精准的匹配和搜索。为了更好地做文字检测，我们提出了一个弱监督学习的精准文本切分的方法。通过投影切分出来的文本与其标注不能精准匹配，要解决的问题是如何进行自动对齐。为此，我们设计了一个弱监督的字符分类器，基于此提出了一种识别指导的古籍文字分割新方法，其核心思想是，针对直方图投影获得的文字过切分结果，基于克努斯-莫里斯-普拉特算法进行改进，进行与文本GT的标注对齐，然后利用一个基于CNN+CTC的识别器去辅助和指导做精准的bounding boxes搜索匹配，高丽藏（TKH）是当时我们能够找到的、公开可以下载的一个数据资源，我们构建了一个TKH数据集来验证提出方法的有效性，整体效果还不错，在高

① 所谓IoU就是交并比，即检测出来的框跟真实的框的交集，除以它们的并集得到的比值。

LoU的情况下能做到90%以上的文字分割精度。我们的方法和当时其他主流方法对比，无论是图片是否足够清晰，均取得了更好的分割精度。

在我们积累了一定的标注数据以后，假设已经有足够多的标注数据，特别是在字符集标注数据充分的情况下，古籍文字检测就不是太难的问题了。比如，只要有几千张不同版式的标注数据集，无论是Faster R-CNN，还是YOLO-v3等，在IoU=0.5的情况下，都能达到90%以上的检测精度，YOLO-v5甚至可达到98%的文字检测精度。尤其是在高精度的检测情况下，YOLO-v5依然可实现很好的性能，总体来讲比Faster R-CNN好。

为了解决高精度古籍的文字检测分割的问题，我们还提出了一个基于强化学习的高精准方法。这个算法的初步思想是利用传统的物体检测方法先做初步检测，然后再用深度强化学习的方法进行精调，目标是希望在高精度的情况下，能够把古籍的文字检测更加精准。强化学习的几个主要的因素，如智能体动作（Action）的设计、惩罚函数、特征表达怎么学习？我们在特征表达中应用了深度学习，Action的设计也非常简单：检测出一个框以后，它会向左右上下、向外向内等去扩展这个框，希望它做出来的框能更加精准，或者它不动就已足够精准。关于强化学习涉及的惩罚函数或奖励函数（Reward Function），我们也针对古籍OCR检测高精度的问题，做了一些定制化的设计。实验结果表明，在引入我们的强化学习的网络（FCPN）之后，传统的目标检测或文字检测的方法都能显著改善古籍文字的检测精度。基于两个公开数据集MTH、TKH进行实验，在高IoU的情况下，提升非常显著。文字检测领域著名的PixelLink或EAST模型，叠加我们的强化学习方法后，甚至可以提升5%—15%左右（IoU=0.8）。这意味着经过强化学习的精调后，它检测出来的框更加精准。

（三）古籍文字识别方法

首先需介绍一下一般的OCR领域主流的识别方法。在深度学习之前，文字识别比较主流的第一类方法是基于分割或过分割的文字识别方法，典型代表如中科院刘成林教授团队2012年TPAMI中提出的分割方法[①]等，近年来仍应用在手写体或古籍文字的识别引擎中。我们团队今年提出了一种新的基于分割的文字识别方法，[②]识别结果比现在主流的CTC及Attention要好。

从2015年以来，中文文字识别领域比较主流的方法大多是无分割的方法。

①Q. Wang, F. Yin, C. Liu, “Handwritten Chinese Text Recognition by Integrating Multiple Contexts,” IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI), vol. 34, no. 8, 2021, pp. 1469-1481.
②D. Peng et al., “Recognition of Handwritten Chinese Text by Segmentation: A Segment-annotation-free Approach,” IEEE Transactions on Multimedia(TMM), 2022, https://doi.org/10.1109/TMM.2022.3146771.

在英文的文字识别领域，十多年前提出的CTC序列建模框架，如LSTM+CTC[①]是做序列识别非常主流的方法，在手写体英文识别当中发挥了非常大的作用。第一次将其成功用来解决场景文字识别问题的是华中科技大学白翔教授团队提的CNN+BLSTM+CTC模型[②]，这个模型是一个非常好的、影响深远的框架，在网上有公开的开源代码。只要有足够多的标注数据，很容易训练出自己的文字识别引擎，包括古籍文本行的识别引擎。

文字识别的第三类别方法是基于注意力机制的文字识别方法，即所谓Attention方法。在OCR领域，近年来代表性方法是ASTER[③]，它把文字图像矫正、特征编码和识别解码输出端到端地结合在一起。基于Attention的文字识别，近几年研究非常多，如SVTR方法在复杂自然场景的文字识别率做得很好。[④]另外，引入语言和语义知识，可以帮助我们更好地做文字识别，中国科技大学团队2021年在CVPR提出一个方法，把语言模型和数学模型很好地联合在一起，端到端地训练。[⑤]语言模型在OCR领域已应用几十年了，但基于Transformer的语言模型和基于Attention机制深度学习网络联合起来端到端训练，效果也做得很好。近年来，也有学者把CTC和Attention方法联合起来使用，利用它们的优点和互补性。[⑥]

下面简单介绍在古籍大藏经中用的文字识别方法，实际上古籍大藏经识别并没那么困难，如果有很好的充足标注数据，并不需要特别复杂的方法。但我们五年前开展这项工作的时候，发现几乎没有公开的数据集，自己标注数据的成本很高。当时我们找到了高丽藏（韩国的大藏经）数据，它的数据量很大，包括16万张图片、5,000多万个字符，但只有图片到文本的篇幅级对齐的标注，并没有

①Alex Graves et al., “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks,” Proceedings of the 23rd international conference on Machine learning(ICML), New York, 2006, https://www.cs.toronto.edu/~graves/icml_2006.pdf.

②B. Shi et al., “An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition,” TPAMI, vol. 39, no. 11, 2017, pp. 2298-2304.

③B. Shi et al., “ASTER: An Attentional Scene Text recognizer with Flexible Rectification,” TPAMI, vol. 41, no. 9, 2019, pp. 287.

④Y. Du et al., “SVTR: Scene Text Recognition with a Single Visual Model,” Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence(IJCAI), 2022, https://doi.org/10.48550/arXiv.2205.00159.

⑤Shancheng Fang et al., “Read Like Humans: Autonomous, Birdirectional and Iterative Language Modeling for Scene Text Recognition,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2021, https://doi.org/10.48550/arXiv.2103.06495.

⑥W. Hu et al., “GTC: Guided Training of CTC Towards Efficient and Accurate Scene Text Recognition,” Proceedings of the AAAI Conference on Artificial Intelligence(AAAI), vol. 34, no. 7, 2020, http://dx.doi.org/10.1609/aaai.v34i07.6735.

能直接训练OCR引擎的检测或文本标注数据。我们就想，基于这样的篇幅级对齐的标注，能不能在没有文字或文本行标注的情况下，利用弱监督学习方法，构建一个不错的文本行或文字识别引擎呢？如果能设计自动切分方法，跟它的文本标签对不齐，也可以解决标注数据的问题。解决标签对不齐的问题，可以用传统的方法，在没有任何标注的情况下，把图片切成一列一列的文本行图片，其中肯定有一些样本是对不齐的，因此我们设计了一个自适应门控机制的训练策略，来减轻对不齐对它的影响。实验表明，如果分割线方法做得比较好，95%以上的情况可实现自动对齐。在此基础上，我们设计了两个简单的识别模型，CNN+CTC和CNN+LSTM+CTC。利用高丽藏的十万余数据，随机挑了75%做训练集、25%做测试集，发现总体效果还是可以，基本上可达到98%以上的识别精度。[①]

我们在古籍识别方面的第二项工作是弱监督古籍文字识别。在没有字符标注、但有文本行标注的情况下，能否与文字切分和文字识别联动，构建出一个不需要精准文字标注数据的文字识别引擎？对此，我们提出了无监督增量自适应学习方法。[②]识别引擎很简单，就是一个卷积神经网络。训练样本是通过弱监督的方式不断地获取标注数据。识别网络可以先用传统的方法标一些数据进行训练，然后通过识别引擎给出的置信度，设计一个基于置信度自动标签机制，这个机制帮助我们将训练集的样本打上标签。我们提出了弱监督门控标签生成的机制，是此工作一个主要的新贡献。在没有用这个方法之前，识别精度大概只有91%；用这个方法后，单字识别精度可达到98%，同时检测效果也有所提升，从86%提到了91%。整体而言，这个方法的识别器、检测器和分割器可以很好地互动。在文档行识别实验中，我们测试了大概八九种大藏经的不同版本，包括《毗卢藏》《思溪藏》《赵城金藏》《乾隆大藏经》《中华大藏经》等，总体置信度还是很准的。

第四个方面，古籍版面分析和端到端的识别。

文字检测或文字识别，包括单字的检测、文本行的检测识别，只要有标注的数据就不是难题，现在深度学习技术已经很成熟，可以解决得很好。但古籍版面分析有时会更难。做OCR的第一步要先进行版面分析，否则无法按正确的阅读顺序进行识别输出。我们团队在古籍版面分析方面做了三个方面的工作：

一是端到端的版面分析、文字检测和识别的框架。针对一张古籍图像，将其版式大的框架检测出来的同时，可不可以将细粒度文字文本框和字符框也检测出

①H. Yang, L. Jin, J. Sun, "Recognition of Chinese Text in Historical Documents with Page-level Annotations," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, https://ieeexplore.ieee.org/document/8583761/.

②Zechen Xie et al., "Weakly Supervised Precise Segmentation of Historical Document Image," *Neurocomputing*, vol. 350, 2019, pp. 271-281.

来？一个模型可以把这三个任务都完成。我们设计了一个深度网络模型，包括版面的分支，文字定位检测的分支及分类的分支，三个分支联合端到端优化。[①]基于多样式的大藏经古籍进行实验，效果也还不错。这项工作的论文大约在2020年发表，与当时一些主流商用系统相比，我们的效果明显要好。无论单列、双列、三个版面、四个版面等，其可视化的版面分析效果基本较好。

第二个方法则基于我们正在做的一项工作来进行介绍。DADSeg是基于Transformer的古籍版面分析新方法，此类模型基于实例分割的思想，利用Transformer的自注意力机制来对要版面分割的位置进行学习和预测，通过一个分割的预测头，经过后处理，可对复杂的古籍版式进行版面分析。这个方法有四个模块：CNN的特征提取、可变形注意力特征融合Transformer block、解码器及后处理模块。在公开的一些古籍版面数据集上实验，如cBad2017、MapSeg2021版数据集，我们的方法效果与目前SOTA的方法相当（cBad数据集）或更好（MapSeg数据集）。基于ICDAR RDCL2015年公开的数据集进行实验，我们的方法比之前的方法明显要好很多；基于笔画级的分割数据集，我们的方法则基本达到了SOTA的效果。

最后介绍我们面向古籍文档的阅读顺序理解方面构建的一个方法。在古籍OCR过程中，能够产生正确阅读顺序的输出是很重要的，否则，如果对一个有四个栏框的古籍图片进行识别，上下栏框的文字可能串成一行，输出完全混乱的结果。因此，我们提出了一个基于图神经网络和Transformer的阅读顺序方法，包括字符阅读顺序和文本行阅读顺序的两个模型。字符阅读顺序是基于图卷积神经网络来构建的，它的每个图的每个字符都构成一个结点，结点特征是由视觉特征和几何特征拼接组成，边特征为两个字符结点的几何关系，然后利用标准的图卷积神经网络推理，从而实现字符阅读顺序的正确输出。

在文本行阅读关系的预测方面，我们基于Transformer构建了一个文本行阅读的方法。这个方法受到EMNLP领域中的LayoutReader方法启发，同样基于Transformer进行设计。评测标准有传统的召回率、调和精度，包括文本行连接关系评价指标ARD等，将其应用于中科院的古籍数据集和我们自己的数据集，根据实验结果，我们的方法比基于启发式的规则明显好。高丽藏图像文本有大字、小字，有大行、小行，其版式实际上很复杂、并不好处理，尤其是如果图片有点背景噪声的干扰，利用启发式规则等常会出错，基于学习的方法能较好地解决这个问题。

①Weihong Ma et al., “Joint Layout Analysis,Character Detection and Recognition for Historical Document Digitization,” ICFHR, 2020, https://doi.org/10.1109/ICFHR2020.2020.00017.

我们最近做的一项工作，是在没有足够充分的数据标注情况下，解决文本的分割、识别和阅读顺序理解同时输出。古籍的版式风格多样，要考虑如何用比较低的成本解决版面分割、识别、预测等问题。我们的方法效果不错，它的优点之一是可以处理弯曲、变形的古籍，如手机拍摄的古籍图片，或有些古籍属于不能下压扫描的珍贵文物，非接触扫描出来的可能弯曲、变形的图片。①

最后给大家分享一下古籍OCR技术的应用。国内专业做古籍OCR的公司不少，如爱如生、书同文等。我们实验室也做过大藏经的古籍OCR引擎，现在开放给大家试用。虽然这个引擎是针对大藏经的数据训练出来的，但针对其他类型的古籍，整体识别效果还是不错的。

下面向大家介绍一个比较专业的古籍数字化的工具平台——“如是古籍数字化生产平台”②，由北京如是人工智能研究院建设，包括OCR、标点、文本对比、切分、纠错等非常好的古籍数字化工具。他们与我们实验室合作，其OCR识别引擎和检测引擎是由我们提供的。平台功能由四个部分组成：一是OCR工具。用户上传图片，平台支持不同版式的版面分析，调用后台的识别引擎就会显示识别结果。针对识别结果，可进行校对和纠错，包括文本纠错和检测框纠错、阅读顺序纠错等多种纠错方式，甚至应用了一些人工智能辅助纠错的技术，具备很好的人机交互方式，可以较好地找到有问题的地方。如果发现文字检测出来的结果有错误，可以进行微调，也可利用工具平台对古籍进行管理，将识别结果以json文件格式导出。二是自动标点。三是标点迁移。四是文本智能比对，如《金刚经》从古到今有很多版本，不同版本存在一定差异，如果能将不同版本间的文本差异自动标识出来，对于古籍文献学及古籍校勘等研究会有很大的帮助。平台支持个人用户注册及团体用户注册，团体用户可以对多人进行管理。

我们实验室研发了“中文古籍文档分析识别演示系统”③，利用一张GPU卡支撑文字检测和识别，还有阅读顺序、版面分析等功能。总体来讲，识别结果的准确率还可以，欢迎大家测试及反馈意见。

我们最近还研发了一个文言文到现代文的翻译引擎“文言文神经机器翻译演示系统”④，它实际上与OCR所应用的部分技术是相通的。从技术上看，古籍的文本行识别是一个序列到序列建模的问题，文言文到白话文的翻译同样是要解决数

①Dezhi Peng et al., “PageNet: Towards End-to-End Weakly Supervised Page-Level Handwritten Chinese Text Recognition,” *International Journal of Computer Vision*, vol. 130, 2022, pp. 2623-2645.

②“如是古籍数字化生产平台”网址：https://lab.tripitakas.net:800/user/login?next=/，海外镜像网址：http://guji.world.rushi-ai.net。

③网址：http://47.101.165.49/textv2/lineRec.html。

④网址：http://region-3.autodl.com:51960/translation。

据问题和模型训练问题。与专业的翻译软件相比，我们的训练集很少，但不少古文翻译得还不错，初步实验结果看起来比百度要好。

比如，《左传》中的一段文字：“初，郑武公娶于申，曰武姜，生庄公及共叔段。庄公寤生，惊姜氏，故名曰‘寤生’，遂恶之。爱共叔段，欲立之。亟请于武公，公弗许。”[①]对我这个理工科出身的人而言，并不好理解，看完翻译的文本后（翻译结果为：“当初，郑武公在申国娶妻，名叫武姜，生了庄公和共叔段两个儿子，庄公降生时是脚先出头后出的，这是难产，使姜氏很惊讶，因此给他取名叫寤生，并且很讨厌他。姜氏喜爱共叔段，想立他为太子，屡次向武公请求，武公不答应。”），就大概知道意思了。我问了一个文科的同学，以五分为满分，这个翻译效果可达到三分，勉强可懂，但比不上专业翻译。其主要问题在于缺乏标注数据，当然将来也可以研究无监督、弱监督或大规模预训练模型的方法解决文白翻译问题。再看一下佛经的翻译，《心经》中“观自在菩萨行深般若波罗蜜多时，照见五蕴皆空，度一切苦厄”[②]的翻译结果为“观世音菩萨用深入的般若智慧进行观照时，照见色蕴、受蕴、想蕴、行蕴、识蕴五蕴都空无实性，以此度脱一切的苦难和灾厄”。“五蕴皆空”的“五蕴”都翻译出来了：色、受、想、行、识，还挺令人惊讶的。对于我们理工科的师生而言，这样的工具可以帮助我们更好地理解古文，大致看懂中国古代流传下来的古籍文献。

今天主要从数据、方法和应用三个层面介绍了古籍OCR相关技术进展情况。展望未来，古籍OCR还有很多问题没有解决。第一，超大类别的古籍识别问题做的还不太好，我们现在的识别引擎最多支持一两万个类别。GB18010-2005数据集有七万多个类别，包括不少异体字、生僻字等，到目前为止还没看到能将其全覆盖的标注数据集。是否有办法先利用一些小样本或合成样本解决这种超大类别的古籍识别问题？第二，古籍图片的修复也是一个有待加强研究的重要问题。如《房山石经》是从石头上拓出来的，所以很多字都没有了，是否有办法通过人工智能技术将这些字进行修复？第三，复杂版式古籍的版面分析。第四，当前的以深度学习为代表的人工智能方法是一个数据驱动的方法，除数据之外，能不能把知识也利用起来？比如，针对海量的古籍，用已经整理好的文本知识或知识图谱，帮助我们更好地解决古籍文档图像的理解、识别等问题，这是值得研究的。第五，通用的古籍OCR的工具平台，国内已经有很多团队在做，包括北京如是人工智能研究院、北京书同文数字化技术有限公司等，开发了非常优秀的古籍数字化工具。另外，我还想特别呼吁一下，OCR领域、数字人文领域、古文献学或

① 阮元校刻：《十三经注疏》，北京：中华书局，2009年，第3724页。
② 藕益智旭撰：《般若波罗蜜多心经释要》，成都：巴蜀书社，2014年，第46页。

古文字学的不同团队能否进行合作，构建、开放更大规模的中文古籍数据集？据我了解，目前开放的可免费访问的中文古籍数据集还不多。相信构建一个大规模的、中文的开放古籍数据集，对促进整个领域的发展有积极的作用，对国家古籍文物的保护与修复，对古籍OCR技术进步、古籍知识图谱构建和信息发现等都很有帮助。我希望不同学科的团队多合作，人工智能、数字人文、古文献学、古文字学的专家学者一道，共同推动古籍OCR事业的发展。

刘永革（安阳师范学院）：甲骨文大数据以及相关识别的处理

我来自安阳师范学院，安阳是甲骨文的故乡，有世界文化遗产——殷墟，我们实验室是甲骨文信息处理教育部重点实验室，也是河南省重点实验室，是国内唯一的一个以甲骨文为研究对象的理工科实验室。我们实验室从2018年开始建设，主要围绕国家重大特殊需求，以服务文化传承为主线，建设具有国际水平的甲骨文大数据平台，采用智能技术服务甲骨学研究，目前主要有四个方向：甲骨文大数据平台、甲骨文识别与字形分析、甲骨文语言计算、甲骨文与殷墟科技考古。

甲骨文是我国最古老的文字。“甲骨文是迄今为止中国发现的年代最早的成熟文字系统，是汉字的源头和中华优秀传统文化的根脉，值得倍加珍视、更好传承发展。”这是2019年习近平总书记致甲骨文发现和研究120周年座谈会贺信中的原话。甲骨文也是最古老的语言。它不但是文字，还是语言，有主语、谓语，有语法。根据此次会议的主题，我感觉甲骨文也是最古老的古籍。但是它不是在纸上，而是在甲骨上。

我要谈的主要有三个方面：第一是介绍我们当前的甲骨文大数据建设情况；第二是甲骨文文献的识别问题；第三是甲骨文的识别问题。

我们建设甲骨文大数据平台的基本建设思路也是从数字化到数据化，再到智能化。“殷契文渊”（http://jgw.aynu.edu.cn）是我们实验室的标志性成果，在甲骨学领军人物、中国社会科学院学部委员宋镇豪教授亲自指导下建设、命名的大型的加工数据平台，也是目前全世界最大的甲骨文数据库。“殷契文渊”里面包括三个子库：字形库、著录库、文献库，是一个免费的开放平台，2019年发布了网页版，2021年发布手机版，所有的资源都可以下载和访问，使用手机号码即可注册。

著录库主要收录研究甲骨文最原始的材料，包括照片、拓片等。书有很多，

目前我们收集了240多种，上线152种。著录库的主要内容是甲骨片的原始数据，如甲骨的拓片、墨本、正反面照片等关于相应甲骨的数据。其次是释文，包括甲骨上的文字及释义，目前我们公布了国内甲骨学界比较认可的四种释文，可供参考使用。再次是重片。现在数据库收录23万片甲骨。但据统计，目前公布的甲骨片实际上为16万片，多出的这几万主要是重片——某些甲骨可能著录在不同文献中，所以就重复了。然后是缀合。某一片甲骨能否和其他甲骨缀合？甲骨片在地下埋藏了三千多年，质地较脆，发掘过程当中或平时运输中等都会造成甲骨的破碎，所以95%的甲骨可能都是碎片，因此缀合非常重要。比如，《甲骨文合集》中的第一片甲骨，可以和《甲骨文合集补编》中的一片甲骨缀合，缀合后文字信息就更多了。我们实验室采用计算机技术进行缀合，从2019年至今，已经缀合出了人类专家还没缀合出来的36组甲骨，并且经专家用实物验证也是正确的。最后是相关文献，即针对某一片甲骨的研究文献，这也是甲骨文专家需要的知识内容。

文献库主要收录研究甲骨文的论文、专著、会议文章等。从语言来说，有英文、法文、日文，当然中文是最多的。它和前面专家讲的古籍相比有很大的不同，关键在于其中包含大量的甲骨文、篆文等古文字及相关图像。版式有横排、竖排，字体有手写、印刷，识别难度较大。我们目前做的一项工作就是把文献库中关于拓片的信息提取出来，并和著录库进行关联，如此文献库、著录库之间就可以联通了。利用文献查拓片，通过拓片查文献。

字形库中的内容以当前甲骨学研究领域公认较权威的四本书（《甲骨文字编》《新甲骨文编》《甲骨文字新编》《甲骨文字形表》）为基础，并对其进行了一定的重新整理，我们应用计算机技术进行字体的分类、部首的合并等，并解决了一些误收、重收等问题，形成了一个新的甲骨文字形表。我们已经把字形表公布在互联网上，而没有出版为纸质书籍，因为书出版之后，就很难再改动。字形表在“殷契文渊”就可以访问，其中包括部首、单字、异形字等。通过单字也可查询其字在哪些甲骨片上出现、在诂林中的释义、字链接的《甲骨文字编》条目等。

我们把三个库建立了关联，集成在“殷契文渊”网络平台。通过网络平台，能够把研究甲骨文需要的文字、图像、卜辞、释文、相关文献在语义层面进行有机的结合，辅助甲骨文学者完成人工难以完成的很多工作。

甲骨文文献识别有哪些意义？主要有三个方面：一是国家的特殊需求。习总书记关于甲骨文曾有三次讲话，强调甲骨文是汉字的源头和中华优秀传统文化的根脉，研究传承甲骨文，有利于增强文化自信。二是甲骨文研究的需要。甲骨研究者不单需要看著录信息，看原始拓片，还要看文献——这个字哪些人研究过，

这片甲骨哪些人研究过。图片的形式形式，不能融入信息化场景，如无法进行检索、信息抽取、构建知识图谱等，这种形式限制了甲骨文研究工作的开展，所以甲骨文献必须数字化。三是释读甲骨文的需要。因为目前甲骨文有4,000多字，其中已释读1,000多字，还有约3,000字尚未解析，甲骨文文献的释读，仍然有很大进展空间，意义非常重大。

当前甲骨文文献识别仍存在一定问题。我利用书同文公司的“i-慧眼OCR”、阿里巴巴公司达摩院的“读光OCR”及古联数字公司的OCR共三个产品，对甲骨文献识别做了实验。我选择了两页甲骨文文献，一页手写竖排，一页印刷横排，分别在三个系统进行测试，看效果如何。

书同文的“i-慧眼OCR”识别效果确实较好。当然，其中两个甲骨字没有判断出来。“i-慧眼OCR”的校对功能还是不错的，凡是识别错的地方，我就用红线划了出来。横排页面中的文字切割得非常好，当然相应的识别结果，仍有出错的。其次是达摩院的“读光OCR”，它的文字切割还可以，但是古文字识别存在问题。我们所说的古籍文字，通常指汉代文字定型之后的古文字。那么这种先秦古文字是不是算古籍文字？“读光OCR”识别印刷横排页面中的甲骨字，肯定不成功，还有一些其他字识别得不对。古联数字公司昨天公布了新系统，我也做了实验，同样对于甲骨文的识别存在一些问题。

那么，就是目前OCR识别软件对甲骨文文献的识别情况，我将甲骨文文献识别主要存的问题总结为四方面：第一是甲骨文字识别不出来。第二是隶定字识别不出来。什么是隶定字？隶定字是古文字专家为了写印或为了说明字的意思，把古文字用现代汉字的部首进行构型而创造的，是古文字研究领域特有的概念。第三是文中的小图片识别不出来。写论文的时候要说明字形，而字形在字库里没有，所以制作了小图片放在文中。甲骨研究文献中存在大量的这类小图片，如何把这些小图片识别出来，是一个重要的问题。四是手写如果不太清楚或有专名线，则不易识别。手写文字不清楚有很多原因，可能手写或印刷时不清楚，或后期扫描不清楚。

其中的甲骨文识别问题则可分为三类：一是甲骨文单字识别；二是甲骨文图识别；三是甲骨图像中甲骨文的识别，即能不能找到并识别出甲骨原片上的甲骨文字。

甲骨文的特点之一是异形体比较多，比如甲骨文字  ，隶定作“  ”，这个字有57个异形体，中间或上面是一个羊，两侧或下方是两只眼或一只眼，可能有的字形只出现过一两次，目前还未被释读出来，但甲骨专家认为这57个异形体指的是同一个字。那到底专家说的对不对？我们可以用信息技术进

行分类。2020年，为了开发手写输入法，我们构建了一个手写数据集，叫做HWOBC。甲骨文字库的字符数量一直在增加，从刚开始的3,881个甲骨文字，到现在已达到四千余个。其次，是我们和华南理工大学共同建设的OBC306数据集。在我们的前期工作中，构建了一个30多万张图片的数据集，即截取拓片上的甲骨文字并对其进行识别。甲骨文存在模糊、背景噪音较大、字形残缺等问题，所以现在识别率不高，只能达到约82%。此外，我们自己建设了一个甲骨检测数据集。这项工作基于我们承接的一个国家级项目，对甲骨文进行检测与识别。首先要从拓片上把文字找到，再做识别。所以为了检测甲骨文字，我们做了一个小的数据集，有八千多张图片。对每个字的位置进行标注，然后用深度学习方法进行训练。

这三个数据集都在“殷契文渊”平台开放，大家如果有兴趣，可以下载做实验。当然这些数据集有很多不完备之处，请大家多提意见，我们再进行修改。

关于甲古文的OCR识别，我们实验室有个梦想，即利用信息技术破译甲骨文。比如这个  字，它的部首大家都认识——中间是“老”字，但“老”两侧有两个“口”字。这是什么字？到现在专家也未释读出来。所以我们就想能不能用信息技术，用大数据、知识图谱等方法，把字给释读出来。

甲骨文作为中国最古老的文字、最古老的语言和最古老的书籍，就像一个有待开发的宝藏，有很多价值可去挖掘，希望感兴趣的专家、学者、学生都可以参与其中。

张弛宜（北京书同文数字化有限公司）：古籍 OCR 工程应用 25 年

每一次研讨，我们在自己软件的功能设计上都会有突破，我们常常是站在高校学术的巨人的肩膀上往前走的。我主要向大家分享25年中OCR在古籍全文数字化工程中的应用。

经常有人问我，书同文的古籍OCR的识别率是多少？我不敢说识别率到底有多少，因为古籍的版面变化太大了，有的只能识别50%，有的甚至能接近100%。我们不可能取一个平均数来说平均。所以每当解答客户提问时，总有一些前置条件，这样去论述才是最准确的。那么，我今天所分享的书同文古籍OCR是有一定前置条件的，第一是在古籍全文数字化工程上使用；第二是针对古籍较规范的版面；第三是古籍内的字迹识别是非特定人规范手写汉字识别；第四，我

谈的古籍OCR是一种泛OCR工程的应用。为什么是泛OCR？因为我今天做的分享，其实严谨来讲有很多技术点不是OCR的范围，但会真正地使用在古籍全文的数字化工程中，为了流程完整性，我也就全部提及：第一是图像的预处理；第二是版面分析，文字检测；第三才到真正的核心，即OCR单字识别——单个字迹的识别；第四，把OCR提供的结果使用在校对流程中，制作的一系列校对工具；第五是通过OCR的算法衍生出来一些由版式的物理关系归结到语义逻辑关系。

书同文古籍OCR，在25年来经历了三个阶段——古籍OCR工程需求的驱动，瓶颈坚守和突破创新。

1980年代，OCR在中国大陆就已经非常广泛了，但古籍OCR在工程上的制作、实用确实是来源于一个项目，也是我们书同文的第一个项目。1997年的文渊阁四库全书全文检索版工程中心，是我们开始的名字。我们最重要的技术突破，是首次将国际编码字符集和OCR技术应用于大型古籍全文数字化工程。

那时的古籍OCR叫做“特定人规范手写”，因为抄缮《四库全书》的人都是钦定的，字迹非常规范。1997年，我们基于清华大学智能技术与系统国家重点实验室马少平教授团队的算法进行了改进。当时为什么要和清华大学合作？其中有一个小故事。张钹院士在2020年的一篇文章中提到：“AI除了洞见，更有沉淀。”1978年清华大学培养了第一届人工智能硕士研究生，我们公司的创始人张轴材先生恰好是六名硕士生之一。因为这样的渊源，他就和当时的人工智能研究室进行合作，开发了基于古籍的OCR技术。在校对生产过程中，很多北京大学、北京师范大学、清华大学、复旦大学的学子，还有各大图书馆的工作人员，都参与了这个项目。

中国人工智能学会马少平教授带领的一些师生做了一个短片，虽然讲的是25年前《四库全书》是如何数字化的，但在那时已奠定了古籍的标准全文数字化流程。今天虽然我们的算法和算力发生了变化，但古籍全文数字化制作流程基本是从这里就有个高度，而且基本形成了行业的标准。1990年代，清华大学提供了人工智能古籍版面分析引擎和OCR识别引擎，北京书同文有限公司制作了全文数据，并研发了全文检索引擎和最终使用端管理端平台，在香港迪志文化出版有限公司投资下，书同文公司历时三年，完成了《四库全书》的全部数字化工作，并由上海人民出版社发行出版。那么这时候的OCR是怎么实现的呢？（以下为马少平老师的视频文字内容）

一、版面分析数字化的第一步是对扫描出来的《四库全书》页面进行分析，将页面分割出一个个汉字。看起来简单，但实际情况要复杂得多。例如，正常字体的正文之间可能会插入一些小字注释，有些相邻的字可能存在空间上的交

叉，还有一些麻烦是编纂、抄缮的官员人为制造出来的——比如，为了拍皇帝的马屁，他们会故意弄一些错别字，让皇帝审查的时候找出来，以彰显皇帝英明博学。同时，为了表示对皇帝的尊敬，他们在书中提到臣子时会用小字，提到皇帝时则用突出的大字。这些都给版面分析带来了很大困难。

二、非线性整形变换。原本的《四库全书》由三千余人抄写而成，虽然整体比较工整，但字体因人而异，差别较大。为了提高识别率，系统首先采用非线性变换的方法对汉字进行归一化处理，使得同一个汉字看起来尽可能是一样的。它们保留了原来的书写特征，但看起来有更好的一致性。

三、古汉字识别。对古汉字的识别采用近邻法，首先提取汉字图片的统计特征，再基于该特征计算马氏距离作为汉字间的相似度。因为古籍中的汉字缺少训练样本，研究人员采用了滚雪球式的增量学习方法。首先建立空的识别字典，每当遇到字典中没有的汉字时，由人工进行标注后加入字典中继续学习。

四、人机协同的确认与校对。《四库全书》包含了近3万个不同的汉字种类，这是一个非常庞大的分类系统，由于训练样本有限，最终的首选识别率大概在95%左右，但是十选识别率可以达到99%以上。为此，研究者设计了一个人机协同的识别确认系统，机器提供汉字的图像以及最优的十选识别结果，由人工选择正确的汉字。经过这样的人机协同校对，抽查结果显示错误率小于万分之一，达到精品出版物的水平。”

有人提出，我们现在能标注的训练集非常少，能不能大家在社会层面一起进行合作？我也非常期待这种合作。我们的项目开创了古籍全文数字化数据处理的制度流程，人文专家和技术专家的合作在25年前到达了制高点，高校学术和工程实践结合，并且具备全球化的视野，当时使用了single data single binary技术。我们期待在25年后的现在，可以实现更大突破，因为时机已经比较成熟了，大家在各方面能够进行多方的合作，进行再一次的飞跃。

从1997年开始，虽然我们一直在做古籍全文数字化相关工作，但我可以很实在地说，这二十年间我们没有完善OCR引擎，也没有突破。而且我们当时把OCR的技术和工具，销售给了一些数据商，甚至当前数据商都已经放弃了这种OCR的核心路径。我们也进行了一些探索，却始终没有突破，原因在于古籍图像有很多不规则的地方，图像纸张退化、水印、纸张污迹、变形、版面复杂、多特定人变成了非特定人等众多因素，造成了OCR的效率非常低，甚至不如人工录入。包括大家熟知的爱如生公司，也逐渐放弃了OCR道路——他们最初建设“中国基本古籍库”时采用了我们的OCR生产线，但是后来也放弃了这一道路，因为OCR确实造成了干扰。尽管如此难，我们还是坚持用OCR做了51个数据库子库。

到了2018年底，借助以神经网络为主的技术，书同文进行了技术路线的探索，在工程实践中实现了突破。我们的产品名字叫做“i-慧眼”，配备一台中档的GPU，目前一天可识别4万叶（筒子叶），也就是约1,400万字。

它可以处理什么？常规的古籍、有侧标古籍和有页眉页脚类的古籍，都可以识别。有人问我，你们现在的OCR引擎能够识别多少汉字？——33,000多万字，及少部分的标点。我们所用到的训练集包括约一千万字迹，同时拥有符合Unicode14.0标准的9万多的大字库支持。所以如果有特别需要，我们会把这另外一个模型提交出来，放到网上供大家使用。

从而我们得出结论，OCR的益处有哪些？第一，OCR确实提高了效率，但最重要的是它提供了一些关联性，包括最终呈现的图文关联性，及校对时原版与文字对应的关联性。第二，富信息（这是我是从北大的俞敬松老师的报告中学到的概念）。以前识别出汉字就可以了，最多还有候选字和识别的距离，而现在我们又增加了什么？从这一页电子数据生命周期开始，到校对过程，最后到全文检索发布截止，所有的校对的痕迹都应该附着在电子信息上，即完整的富信息。这其中包括识别的正确字、物理坐标位置、候选字和评分标准，还有校对流程中的历史痕迹，以及脱机信息。为什么要特别提脱机信息？是因为我们除了在线上校对以外，还有线下校对。俞敬松老师曾讲，他看到堆在办公室里面的一沓一沓、一摞一摞的《儒藏》校对文稿，觉得非常可惜，因为这些到最后都是留不下来的。而我们则在生产中将这些“校对文稿”做了关联，即脱机的校对信息。

我们在OCR过程中非常好地运用到了置信度，以及校对工具。我们因为是企业，而非学术单位，所以我们的学术理论力量、算法力量一定不如高校，但是又不能耽误工程进度，怎么办？要合理化地规避一些OCR效率不高的地方。例如有的版面，我们做的时候可以分割开，但是会涉及页眉页脚。因为某一个操作，如果应用在一页、两页、十页、一百页的时候，分割是没有问题的，只需确认它的全文检索的信息没有问题。但如果在工程上应用的话，校对员的工作量会有几何级的增长。我们会有预处理的工具，叫做Scan Taylor，这是开源的，做预处理非常方便，是可以不用OCR算法的。比如族谱，虽然把每个字都识出来了，但是并不知道上下文的关系，那么如何在目前没有语义算法的情况下将其连接起来？我们用了一些软件纯代码的方式进行连接。这就是在工程上应用的，有时候要合理地规避一些OCR解决不了的问题，就不能跟算法较劲。那么在浏览端怎么用到OCR？大家可以看分析出来的数据，除了图像和文字页页对应以外，还可以列列对应，字字对应，而且同时可在图像上标记出检索词的位置。

我们把自动识别和校对工具分开，分别叫做“i-慧眼”和“点字成金”。“i-

慧眼”是轻量级的识别校对工具，真正应用在工程上的是“点字成金”。除了恢复版面外，很多古籍有一些很细节的东西，例如页脊、列间的标点、大小字、专名线等都需要识别出来，这些适应用于工程上识别使用。

“i-慧眼”可以单页识别，也可以批量识别。它把每个字迹都切割出来了，可以选择是否保留页脊、标点。列对列的校对方式，保证视线转移非常少。我们使用评分过滤及置信度方式过滤一些错误，通过这种方式我们掌握了一套置信度和评分标准，以及两选之间的距离之差的方差，最后得出了一些统计规律。在校对中常用到的是重点校对，基本可以节省50%，甚至70%—80%的工作量，因为提高效率与减少校对工作量直接挂钩。排印本的文本是有侧标的，一定也要都识别出来。系统不采用NLP的方式，而是保留文本本身带有的信息，能够很快把专名识别出来。

“点字成金”是校对的众包系统，可实现版面分析、列对列的聚类校对、图文校对、重点校对、双校对策略及疑难字、异体字校对。针对某些输入困难的情况，例如有一些生僻字、拼字方法，系统配备手写输入方式。我们的现在的字库支持ISO10646标准和2021年9月公布的Unicode14.0标准，一共是92,853个字，大家可以在我们的网站免费下载。同时，我们有很多的输入规范，提示异写或微小笔形差异，也是公开可下载的。在认同规范中包括简繁不认同、结构不同不认同，源字集分离不认同，这是张轴材先生早期根据国际标准、工作经验等所做的总结，相关说明囊括在一个重要的文档[①]，大家可以参考。

“时有所学，日有寸进。保持热爱，奔赴山海。”古籍数字化，虽然是非常狭窄的行业，但是能够助力增强我们的文化自信，同时囊括很多交叉学科，现在是非常好的机遇，希望广大的师生能热情投入到这一领域。最后分享给大家我们三个较为重要的网站：书同文i-慧眼OCR（ocr.unihan.com.cn）、书同文古籍数据库（guji.unihan.com.cn）和书同文汉字网（hanzi.unihan.com.cn）。

① 文档访问网址：http://www.doc88.com/p-690306085410.html。

附　录

讨论和交流：

问题1：地方志属于古籍文献中内容最丰富的组成部分之一，当前这一领域的使用现状和发展前景如何？

张弛宜：近年来国家地方志编纂办公室投入了大量的研究经费，并且集中各个地方的地方志编纂委员会、编纂办公室力量与党史办公室合作。在地方志的古籍数字化方面，国家图书馆和爱如生似乎已经建设了很多，我们也做了一些。但是对于地方志内容的学术研究，我就不是专家了。

问题2：是否考虑过如何利用智能方法进行纠错？

刘成林：因为没有参与实际的古籍数字化应用的工作，我的回答其实只是设想。从一般做文档处理的角度来说，到了最后自动识别阶段，总是不可避免的会有少量的切分和识别错误。现在有些应用可能不太在乎少量错误，也继续在用。但是对于古籍的数字化或出版工作而言，是有文字质量要求的。所以人要进行校准，首先得发现错误识别或错误切分在哪个地方。总不能把识别的结果从头到尾再仔细去看一遍，那样还不如重新录入一遍，二者可能花的时间差不多。所以，自动识别系统关键要把可能会错的，也就置信度较低的字符，自动标出来。从人工智能、模式识别、机器学习的角度来说，将模式的置信度准确地标记出来是比较困难的问题，可以说是人工智能领域的前沿研究问题。过去有一些研究能够部分地解决这类问题，能得到差不多的置信度，但并不是很准确。所以，从研究角度来看，这方面还是需下很大的功夫的。另外，过去一般怎么做？比如没有置信度估计的话，假定识别正确率为99%，那么1000个字里就有10个字是错的。把这10个错字标记出来的话，过去一般把分类器输出的置信度，如卷积神经网络输出概率的最大值设置在一定阈值，如果最大概率小于阈值则拒识。那么，不可避免地会把一些识别正确的字也拒识，所以如果取不同阈值的话，比如原本错误率是10%，拒识5%，最后整体的错误率能够下降0.1%。如果想把错误降到零，可能需要拒识10%、甚至20%，这样标出来的存疑字就太多了。如果能够实现比如拒识2%，错误率就能降到0，也即原来错误率1%，再把拒识字的2%标记存疑，这样的话，我觉得会对人工非常友好。

问题3：从中文古籍OCR数字化整体发展角度来看，学术界和企业界怎样共同建设和发展领域？是否有相关的规划或数据资源的共享和共建？

刘成林：我们希望跟古籍出版业、古籍数字化行业都去合作。因为实际上如书同文公司是搞应用的，掌握非常多的数据资源，那么过去建设过的数字化典籍，如果图像已经保存下来，是不是可以把那些标记过的数据部分公开，开源并与研究界共享？

张弛宜：我们也有这样的打算，会逐步地开放一些资源。我们也需要与一些高校合作，而且当时与清华大学合作较为成功，我们也基于清华大学的算法，进行了调整和训练，并建设了数据集。这样的合作是多赢的，我们到会逐步开放训练集。

问题4：刘永革老师长期做甲骨文数据资源方面的工作，在数据开放上持什么观点？

刘永革：我们实验室属于研究机构，在研究过程中，只要有成型的数据集，马上就公开了，目前在“殷契文渊”平台已有三个数据集。关于数据资源，我们的思想是开放的，愿意和大家合作。我们掌握一手数据，但在算法上需要刘老师、金老师这样的专家帮助我们。把数据做好，也是我们实验室的责任。

问题5：古籍翻译给谁用？如何保证句读的准确？什么时候正式发布？

金连文：正如我的报告所言，首先，我们今天利用的数据还不是特别多，所以当前只是第一步。当然，对于专业的文科老师而言，可能古籍翻译工具的参考价值没有特别大。但是对于我们非文科专业，尤其是很多希望看懂一些古代的书籍，包括小说的人来说，是有参考价值的。对于某一个专业，比如要读军事、医学、佛经等，那么这样一个工具至少可以辅助大家阅读理解。翻译要达到信、达、雅标准，虽然我们的工具距离雅还差得很远，达也还不够，但能够达到百分之七八十的准确率，对于帮助我们理解古文，已经迈出了第一步。我也呼吁，如果将来有更多的标注数据，也希望有专家、学者、机构与我们合作。如果能够有几百万甚至千万级的标注数据，古籍翻译问题是能够解决好的。现在瓶颈在于数据短缺。那么如何解决？我们希望更多社会力量贡献数据集，我们在界面特意留了纠错按钮，用户如果发现我们翻译的不对，就可以进行修正。将来如果有资源、有条件的话，我们希望把它做成用户登录的界面，纠错达到一定数量，甚至可以考虑发点报酬。

问题6：陕西省文管会已经出版了何汉文先生的《古文字探研》，从考古学和文字学结合的角度研究甲骨文，不知道有没有采集到相关的信息？

刘永革：我们实验室的目标是采集所有研究甲骨的书籍，但是因为毕竟它是冷门绝学，所以有些文献资料还未进入视野。如果能够买到何汉文先生的《古文字探研》，我们一定会进行采集。

问题7：在人工智能时代，古籍OCR领域，包括古籍、尤其是古籍文字识别辅助的古籍数字化领域，还有哪些特别值得我们关注，或者没有解决，或者很少人关注的问题值得我们去研究？

刘成林：这个问题很大，其实我觉得古籍OCR或者文档数字化过程中所有的技术环节都有值得研究的地方。图像预处理、版面识别、后处理、交互、语义理解、大数据挖掘，现在的技术都有不足。但同时，为什么古籍的OCR研究相较于现在一般的文档识别或者人工智能领域其他的技术，比如互联网的搜索，发展比较慢？第一，因为做古籍OCR或古籍数字化的经济效益不明显，所以到现在为止，没能够吸引大量的研究者做这个工作。可能是因为在这一领域的投入相对少了一点，但现在国家对文化、文化工程、数字文化、数据库的建设等方面越来越重视。第二，从学术界的角度来说，它存在大量的技术问题。我觉得，做古籍数字化，大家不用担心不好写文章，同时能够帮助申请国家的资助项目，对学术界来说就有吸引力了。对于企业界而言，吸引力可能不太容易体现，但我相信国家也会越来越重视，越来越增加这方面的投资，所以我对其前景还是非常看好的。

张弛宜：作为工程实践单位，其实我们最关注的是置信度问题，它能够减少非常大的工作量。但是我们现在根据置信度和其他参数，只能做到十万分之一的识别错误率。但这基于筛选留下的样本较多，没看的样本有十万分之一错误率，而印刷体的识别错误率甚至可以达到百万分之一、千万分之一。所以我觉得就古籍数字化而言，降低工作量是最需要突破的事情。

王军：无论是学界，还是业界，一方面存在技术的难题，一方面也存在企业的生存问题，但是我们都在共同促进文化的传承。我想其实我们现在做的工作，包括社会的发展，都是在传承自己的文化。刘成林老师讲得特别好，他的展望是有预见性的。这一领域长时间没有得到充分发展，很重要的原因就在于投入不足。但现在有较好的时间窗口，虽然各个层面仍存在困难，但是从最近国家的几次发文来看，以中共中央办公厅、国务院办公厅联合发布的《关于推进新时代

古籍工作的意见》为例，显示出国家对于古籍相关工作的重视。另一方面，我们在技术上已经比较成熟了，已有比较多的经验和技术积累。更重要的是，现在一些大型企业也对这个方向非常热心。将这几方面的因素整合在一起，我们相信这一领域会有较好的发展，也会对古籍的传承有较大推动。我们北京大学数字人文研究中心希望建立一个促进领域内各方面专家和各方面力量沟通的平台，希望能够发挥一些积极作用，使大家共同合作，更好地携手推动领域发展，产出更多成果，培养更多人才。

整理者：王胤斐 / 北京大学外国语学院
朱　慧 / 四川师范大学文学院
韩静雯 / 武汉大学文学院
徐　璇 / 清华大学《数字人文》编辑部

（编辑：桑海）