运筹学与控制论

DOI:10.11721/cqnuj20240209

融入单元格结构信息的表格抽取方法*

乔　岩[1]，吴至友[1]，高　桓[2]，段旭祥[3]
(1. 重庆师范大学 数学科学学院，重庆 401331；2. 英特尔边缘智能联合研究院，南京 211135；
3. 重庆大学 数学与统计学院，重庆 401331)

摘要:现有的端到端方法和基于预训练模型的方法在训练过程中未有效利用表格单元格的结构信息,从而影响了表格文本在模型中的向量表示和最终的语义信息抽取精确率;因此提出了进一步利用单元格结构信息来改进光学字符识别效果的端到端方法和增加单元格序列预测任务的预训练方法。实验结果显示改进后的 2 种方法在表格语义信息抽取任务中取得了更好的效果,F1 值分别提升了 0.204 6 和 0.017 6。改进后的方法加强了单元格结构信息在表格中的重要性,提高了表格语义信息抽取的精确率。

关键词:表格信息抽取;单元格结构信息;表格识别算法;单元格区域识别
中图分类号:O224;TP391.1　　**文献标志码**:A　　**文章编号**:1672-6693(2024)02-0137-08

随着信息时代的到来,大量的信息以表格文档的形式呈现。作为数据信息的主要载体,表格文档在生产和生活中扮演着日益重要的角色。不同类型文档的具体格式可能有所不同,但通常以自然语言的形式呈现信息,包括纯文本、多栏布局和各种表格表单(图 1)。表格文档存储信息具有集中的信息表达、精炼、清晰、规范等特点,但人工提取表格数据耗时较多且操作成本较高;因此如何更高效准确地提取表格信息成为十分重要的研究课题。

现有的表格信息抽取方法主要分为 2 类。一类是传统的端到端方法,例如 Trie[1]、LES[2] 等算法;该类方法需要使用光学字符识别(optical character recognition,OCR)[3] 工具对图片进行识别,它的特点是不需要提供图片文本识别标签即可完成信息抽取,而在实际情况下大多数文档表格均未提供文本识别标签(图 2)。但是,由于大多数文档整体结构和布局较复杂,因此该类方法使用 OCR 来处理表格单元格区域时不够准确,导致最终语义信息提取[4] 精确率较低。另一类是基于预训练模型的方法,例如最新的 LayoutLMv3[5]、StrucTexT[6] 等算法;该类方法是在已经提供表格文本识别标签的情况下进行语义信息抽取,故该类方法相比于第 1 类方法而言对一般语义信息抽取的精确率更高。值得注意的是,关于文本、布局和图像的多模态预训练模型 LayoutLMv3 在表格理解任务中取得了 SOTA(state of the art)性能。不过 LayoutLMv3 也存在一些不足:虽然该模型在输入阶段引入了单元格结构信息,但在预训练阶段并没有针对单元格结构信息进行预训练,从而影响了最终表格语义信息提取的精确率。

近些年来,越来越多的学者开始将多模态预训练模型中的输入信息如文本、布局和视觉特征等进行融合,并对多模态信息进行处理,进而提升表格信息抽取的精确率。在预训练模型中主要有 3 类方法:第 1 类方法是基于网格(grid)的方法,主要关注于融合图像层面的多模态信息。在这类方法中,文本通常以字符粒度表示,并且文本与结构信息的嵌入方式相对简单,例如 Chargrid[7] 等算法。第 1 类方法主要用于对图像层面结构的表格信息抽取,在抽取视觉层面的有关信息上仍有较大困难,并且在模型输入阶段未考虑单元格结构信息,而是将每个文档页面编码为二维字符网格来实现。第 2 类方法主要基于图卷积神经网络(graph convolutional network,GCN),试图学习图像和文字之间的结构信息,以解决开集(训练集中没有见过的模板)信息抽取的问题,例如包括 GCN[8] 等算法。第 3 类是基于 Token(字符)的方法,此类方法借鉴自然语言处理中的 BERT(bidirectional

* **收稿日期**:2023-12-01　**修回日期**:2024-03-06　**网络出版时间**:2024-05-13T17:03
资助项目:国家自然科学基金面上项目(No. 12371258)
第一作者简介:乔岩,男,研究方向为最优化理论与算法,E-mail:2021110510084@stu.cqnu.edu.cn;**通信作者**:高桓,男,研究员,博士,E-mail:gaohuan@inchitech.com
网络出版地址:https://link.cnki.net/urlid/50.1165.N.20240513.0932.008

encoder representation from transformers)[9]等方法的启发,将位置、视觉等特征信息共同编码到多模态模型中,并且在大规模数据集上进行预训练,从而在下游任务中仅需要少量的标注数据即可获得良好的效果,例如LayoutLM[10]、LayoutLMv2[11]、LayoutXLM[12]、LayoutLMv3、StrucTexT等算法。第2类和第3类方法克服了第1类方法中文本与结构信息的嵌入方式较为简单的问题,在一定程度上提高了信息抽取的精确率。

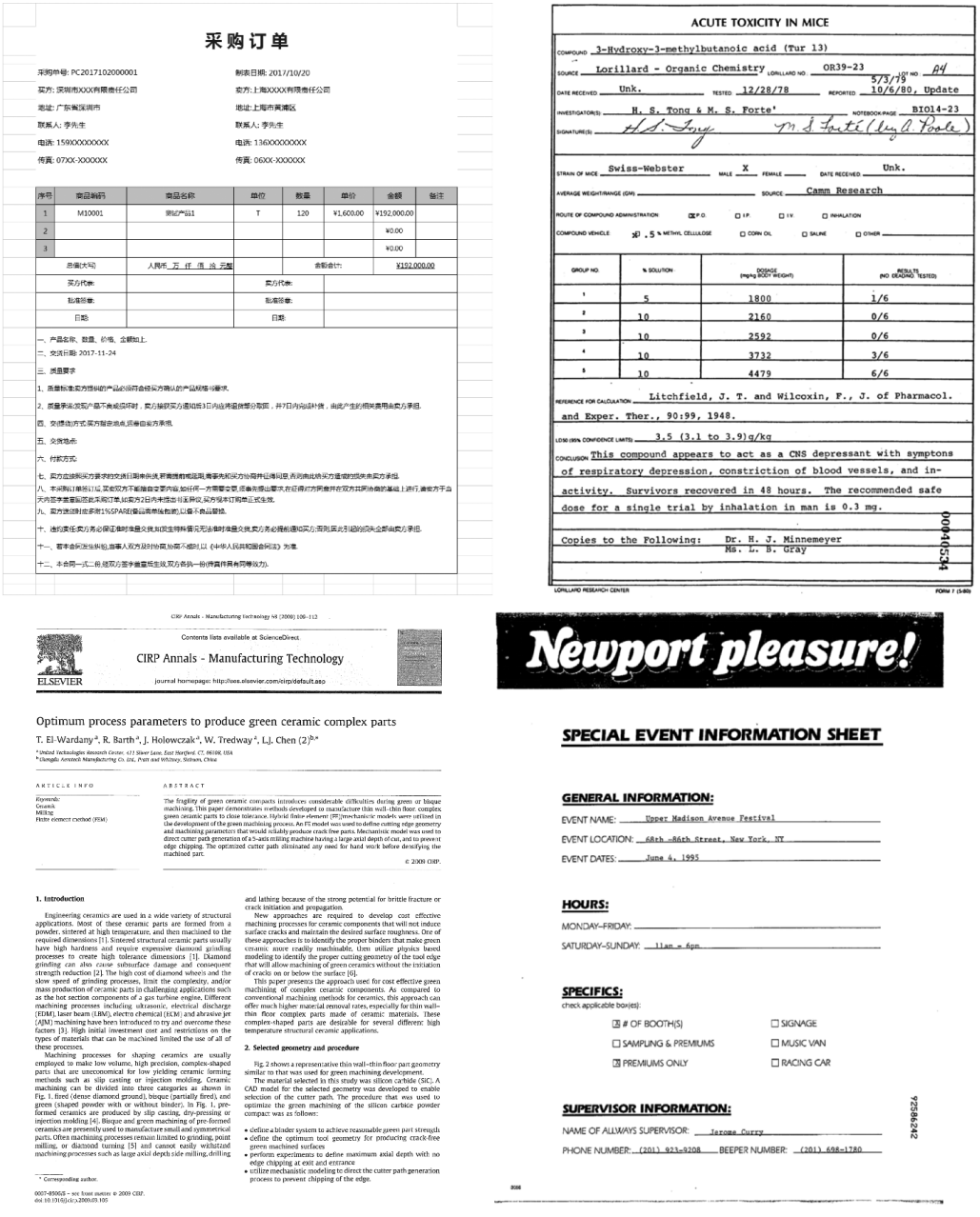

图1 不同布局和格式的文档图像

Fig. 1 Document images in different layouts and formats

受上述方法及StrucTexT的启发,本研究认识到单元格信息对于预训练模型进行表格识别具有非常关键的作用;此外,受StructuralLM[13]的启发,本研究也认识到针对单元格的预训练任务对模型也同样重要。因此,对于传统的端到端表格信息抽取方法中存在的问题,本研究提出了基于单元格结构信息的表格识别算法,通过进一步利用单元格结构信息来优化单元格识别区域来提升端到端语义信息抽取的精确率;对于预训练模型LayoutLMv3中存在的问题,本研究在该模型现有预训练任务的基础上,提出了新的预训练任务——单元格序列预测(cell sequence prediction,CSP),通过该任务进一步利用单元格结构信息优化文本在LayoutLMv3中的向量表示来提升最终的信息预测效果。本研究提出的方法均融入了单元格结构信息,并且充分利用单元格信息的区域特征(语义相同的文本所对应的单元格坐标位置也相同)来提取语义信息,从而使表格信息的提取更加高效和

准确。

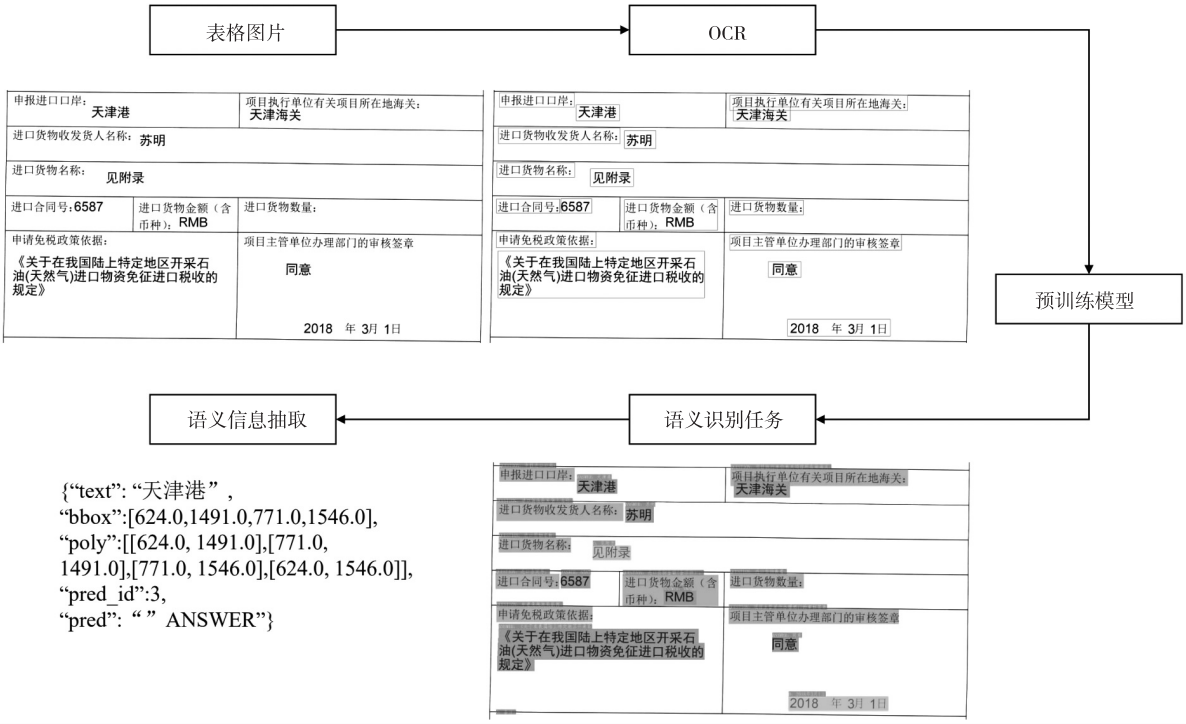

图 2　语义信息抽取流程

Fig. 2　Semantic information extraction workflow

1　基于单元格结构的表格识别算法

在大多数情况下,文档中文字的位置关系蕴含着丰富的语义信息。以图 1 中的采购订单为例,表单通常以键值对(key-value pair)的形式展示(例如"联系人:李先生")。通常情况下,键值对的排布通常是左右或者上下形式,并且存在特殊的类型关系。类似地,在表格文档中,表格中的文字通常以网格状排列,并且表头一般出现在第 1 列或第 1 行。通过预训练,这些与文本天然对齐的位置信息可以为下游的信息抽取任务提供更加丰富的语义信息。

语义实体识别(semantic entity recognition,SER)[14]任务是表单理解中最关键的任务之一,该任务可根据检测到的键值在表格中的位置进行分类,主要分为以下几类实体:标题(header)、问题(question)、答案(answer)和其他(other)。具体来说对于表格 $\boldsymbol{D}$,包含离散的 Token 集合 $t=\{t_0,t_1,t_2,\cdots,t_n\}$,其中 $t_i=\{\omega_i,(x_0^i,y_0^i,x_1^i,y_1^i)\}$, $i=0,1,\cdots,n$, 包含文本内容 ω_i 及它的矩形坐标$(x_0^i,y_0^i,x_1^i,y_1^i)$,模型需要从 Token 序列中抽取出已定义的语义实体。记 $\boldsymbol{C}=\{c_0,c_1,c_2,\cdots,c_m\}$,$m\in[0,n]$为实体类别,则该表格语义实体抽取任务可定义为$(\boldsymbol{D},\boldsymbol{C})\rightarrow\boldsymbol{\varepsilon}=\{(\{t_0^0,\cdots,t_0^{b_0}\},c_0),\cdots,(\{t_m^0,\cdots,t_m^{b_m}\},c_m)\}$,其中 $m\in[0,n]$,$t_m^{b_m}$ 表示第 m 个 Token 集合 t_m 中第 b_m 个字符。

本研究针对 SER 任务下的信息抽取进行了深入分析,并发现主要问题是在使用 OCR 对单元格区域(具体是指每个键值对所在表格的矩形区域)进行识别时不够准确,导致最终的语义信息抽取精确率较低。例如在图 2 中"进口合同号:6587"被错误地识别为 1 个单元格区域,而正确的应该是将它识别为"进口合同号:"和"6587"2 个单元格区域。为解决上述问题,本研究提出一种基于单元格结构的表格识别算法(图 3),它的具体步骤如下:

步骤 1,利用 OCR 对包含多种类型表单的数据集进行识别。该过程得到包含文字和单元格坐标的表格信息文件。进一步分析发现,OCR 结果中的单元格区域不准确,具体表现为单元格漏识别、单元格识别区域过大或过小等情况。

步骤 2,利用 StrucTexT 预测单元格区域坐标。与一般预训练模型不同的是,StrucTexT 在视觉输入层面引入了单元格结构信息,并且通过预训练任务进一步加强了对单元格结构信息的训练,能够较好地预测文本所在

单元格区域坐标。

步骤3,利用StrucTexT预测得到的单元格信息对通过OCR得到的表格信息进行修正。该过程先修正错误的单元格坐标区域再修正对应单元格的文本信息,进而获得更准确的单元格信息和表格信息。

步骤4,利用LayoutXLM进行训练。该模型是一种用于多语言文档理解的多模态预训练模型,可用于弥合视觉丰富文档理解的语言障碍,主要接受来自文本、布局、图像等3种不同模态的信息。这些信息被编码为文本向量和图像向量并共同组成输入向量。输入向量由一个具有空间自注意力机制的多模态转化器进行编码。最后,输出的上下文表示可以被用于以下特定任务层。

步骤5,进行SER任务,对分类得到的语义预测结果进行评估,计算预测指标。

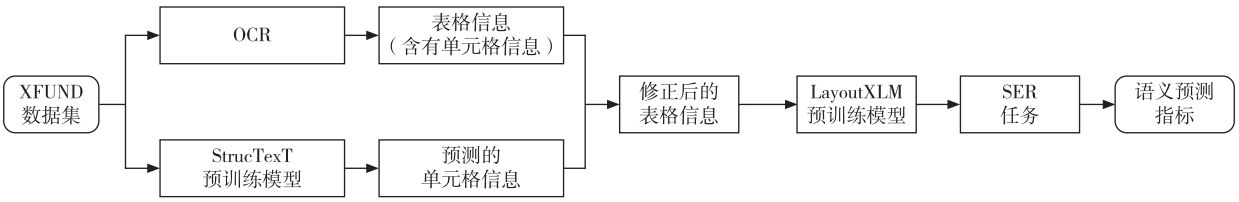

图3 表格识别算法

Fig. 3 Table recognition algorithm

2 LayoutLMv3-CSP

多模态预训练模型LayoutLMv3通过利用Transformer[15]架构学习视觉和文本信息之间的跨模态交互,在预训练阶段整合文本和图像信息。为了克服文本和图像模态在预训练任务中的差异,促进多模态表征学习,LayoutLMv3提出了具有统一的文本和图像掩码任务:掩码语言建模(masked language modeling,MLM)和掩码图像建模(masked image modeling,MIM)。同时,受DALL-E[16]启发,LayoutLMv3还提出了词块对齐任务(word-patch alignment,WPA)用于预测词对应的图像块是否被掩盖。此外,本研究发现StructuralLM在预训练阶段引入了单元格位置预测(cell position prediction)任务以预测单元格在文档中的坐标位置。因此,本研究提出了一个新的预训练任务即CSP,将它用于进一步加强单元格信息在模型中的重要性,并且在最新的LayoutLMv3中进行了多任务扩展。

2.1 模型结构

LayoutLMv3的模型框架如图4所示,它接受的文本、布局、图像等3种不同模态的信息分别被编码为词向量(word embedding)、位置向量(position embedding)和图像向量(patch embedding)。其中,词和位置向量被连接起来,然后加上图像向量,得到模型的输入向量。输入向量由多模态转化器进行编码,该转化器具有空间自注意力机制(self-attention mechanism)[17],可以更好地学习序列中输入的相对或绝对位置关系。

2.1.1 文本向量

文本向量是词向量和位置向量的组合。词向量是通过学习词在上下文中的语义关系并将词表示为低维的实值向量;位置向量则是将序列中的每个元素(例如词或字符)编码为1个向量,以便模型可以捕捉到序列中元素的相对位置信息。位置向量包括1D position和2D position,其中1D position为文本序列内的标记索引,2D position为文本序列对应的单元格区域坐标(矩形边界框坐标)。LayoutLM和LayoutLMv2这2个模型采用字级布局位置,每个字都有自己的位置,本研究则采用单元格级布局位置,因为相同单元格里的词通常表达相同的语义,所以相同单元格中的词共享相同的2D position。

2.1.2 图像向量

目前的多模态模型大部分依赖于卷积神经网络(convolutional neural networks,CNN)或Faster R-CNN检测器来提取图像视觉特征,这需要一定的计算量或需要区域监督。受ViLT[18]的启发,LayoutLMv3在将文档图像输入到Transfomer之前,用图像块的线性投影特征表示文档图像。LayoutLMv3是Document AI中第1个不依赖CNN提取图像特征的多模态模型,这对于Document AI模型减少参数或去除复杂的预处理步骤至关重要。本研究同样也采用LayoutLMv3的图像向量结构。

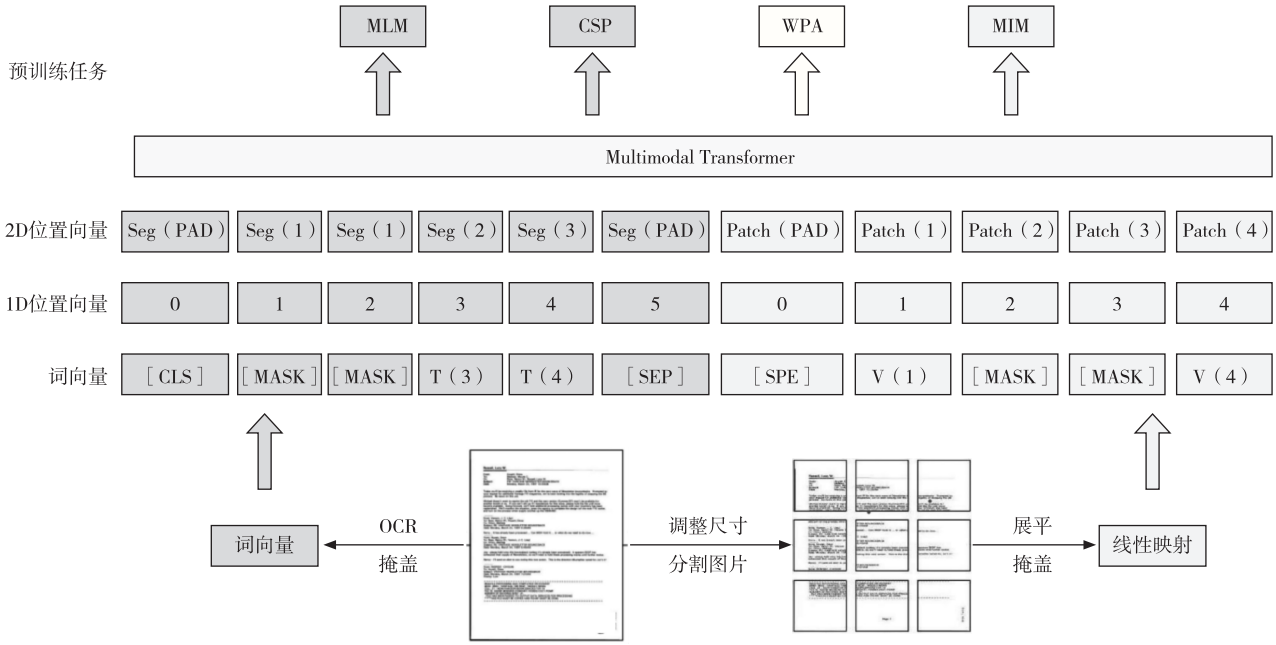

图 4　LayoutLMv3-CSP 结构

Fig. 4　LayoutLMv3-CSP architecture

2.2　CSP

LayoutLMv3 的预训练任务在对视觉丰富的文档进行建模时显示出有效性，因此本研究很自然地将这个预训练框架调整为多任务文档预训练。按照跨模态对齐的思路，本研究的文档理解预训练框架包含 4 个预训练任务，分别是 CSP、MLM、MIM 和 WPA。

尽管 LayoutLMv3 的前几个预训练任务学习了文本和图像之间的联系，但它没有学习单元格位置和文本之间的对齐关系。为此本研究提出了一个 CSP 任务来学习单元格在文本中的序列。具体而言，每个文本都对应 1 个单元格，而语义相同的文本所对应的单元格坐标位置也相同，即具有相同语义的文本对应的单元格序列是一致的。本研究选择未被掩盖的文本，并掩盖它的序列位置，再通过该任务预测文本所在单元格的序列来加强单元格位置和文本之间的对齐关系。预训练任务根据图像标记和文本标记的上下文信息来进行，这 2 个标记分别定义为 $X^{M'}$、$Y^{L'}$（M'和L'代表 Token 被掩盖的序列位置），从而使得下面的交叉熵损失最小化，即预测正确的单元格序列（ω_l）的对数尽可能大，具体公式如下：

$$L_{\mathrm{CSP}}(\theta)=-\sum_{l=1}^{L-L'}\log p_\theta(\omega_l \mid X^{M'},Y^{L'}),$$

其中：p_θ 表示概率函数，θ 表示 Transformer 的参数，$L-L'$表示未被掩盖的文本标记的数量。

3　实验结果及分析

3.1　实验数据集及实验环境

在本研究中，对模型进行评估的实验数据集为 XFUND 数据集（https://github.com/doc-analysis/XFUND/releases/tag/v1.0）。该数据集是一个用于关键信息抽取任务的多语言数据集，包含 7 种不同语言的表单数据，并且所有数据都进行了键值对形式的标注。每种语言的数据集都包含 199 张表单数据（分为 149 张训练集和 50 张测试集），共有 563 个标题、4 343 个问题、3 629 个答案和 1 208 个其他实体，一共包括 9 743 个语义实体。实验操作系统为 Ubuntu18.04.6LTS，内存为 125.5 GiB，中央处理器为 Intel Xeon（R）W-2255CPU@3.70 GHz×20，显卡为 NVIDIA GeForce RTX 3090/PCIe/SSE2，使用 Python 中的 Pytorch 实现代码。

3.2　表格识别算法实验结果

将 XFUND 数据集应用于 StrucTexT 中进行训练，得到了单元格信息的预测结果。为了判断单元格区域是否识别准确，通常将交并比（intersection over union）值设置为 0.5，即预测单元格矩形区域与正确单元格矩形区域的重叠面积超过 50% 即认为是预测正确的。本研究使用精确率（precision）、召回率（recall）和 F1 值（F1-

score)作为评价指标。根据表1可知,在应用StrucTexT后,单元格识别精确率、召回率和F1值分别提升了0.175 4、0.178 1和0.176 8,表明StrucTexT对于预测XFUND数据集的单元格信息具有良好的效果,并为后续的语义信息抽取实验奠定了基础。

将修正得到的表格信息文件与XFUND数据集输入到LayoutXLM中进行训练,并对训练结果进行分类和评估。由于未提供XFUND数据集标签,实验通过OCR获得数据集标签再进行语义识别任务,因此语义预测精确率整体较低。与传统的仅依靠OCR进行端到端信息抽取方法有所不同的是,本研究利用表格识别算法进行OCR得到XFUND数据集标签,然后进行SER任务得到语义预测结果,使得端到端语义信息抽取的精确率、召回率和F1值分别提升了0.205 6、0.203 7和0.204 6(表2)。上述结果表明本研究提出的算法在一定程度上能够提高端到端信息抽取的精确率和其他指标,也验证了单元格结构信息的重要性和该算法的有效性。

表1 单元格识别结果对比

Tab.1 Comparison of cell recognition results

方法	精确率	召回率	F1值
传统方法(OCR)	0.748 7	0.714 2	0.731 1
本研究方法(OCR+StrucTexT)	0.924 1	0.892 3	0.907 9

表2 语义预测结果对比

Tab.2 Comparison of semantic prediction results

方法	精确率	召回率	F1值
传统方法(OCR)	0.447 4	0.426 8	0.436 9
本研究方法(OCR+StrucTexT)	0.653 0	0.630 5	0.641 5

3.3 LayoutLMv3-CSP实验结果

本研究在语义实体任务中将新的预训练任务与LayoutLMv3自身的3个预训练任务相结合,并选择了XFUND数据集作为下游任务来评估LayoutLMv3-CSP的性能,实验结果评价指标仍为精确率、召回率和F1值。表3显示,LayoutLMv3-CSP在各项评价指标上的表现均优于LayoutLM、LayoutLMv2、InfoXLM、LayoutLMv3等对比算法,也说明在文档图像理解任务中对单元格信息进行预训练具有一定优势。此外,将增加了CSP预训练任务的LayoutLMv3-CSP与未增加上述任务的LayoutLMv3相比,前者语义信息抽取的精确率、召回率和F1值分别提升了2.81百分点、0.56百分点和1.76百分点。因此,本研究提出的LayoutLMv3-CSP在文档图像理解任务上相较于未改进LayoutLMv3具有明显优势。

表3 LayoutLMv3-CSP实验结果

Tab.3 LayoutLMv3-CSP experiment results

编号	模型	精确率	召回率	F1值	编号	模型	精确率	召回率	F1值
1	LayoutLM	0.795 3	0.844 0	0.818 9	5	InfoXLM	0.829 5	0.923 8	0.874 1
2	XLM-RoBERTa	0.800 3	0.892 6	0.843 9	6	LayoutLMv3	0.847 0	0.919 1	0.881 6
3	LayoutLMv2	0.822 4	0.892 7	0.856 1	7	LayoutLMv3-CSP	0.875 1	0.924 7	0.899 2
4	LayoutXLM	0.813 7	0.908 6	0.858 6					

进一步采用K折交叉验证LayoutLMv3-CSP实验结果的可靠性,具体步骤为:第1步,不重复抽样将原始数据随机分为5份;第2步,每次挑选其中1份作为测试集,剩余4份作为训练集用于模型训练;第3步,重复第2步5次,每个子集都有1次作为测试集而其余子集作为训练集,在每个训练集上训练后得到1个模型,用这个模型在相应测试集上测试,计算并保存模型的评估指标。

表4显示:LayoutLMv3-CSP实验精确率均在0.86左右,平均精确率为0.865 3;召回率均在0.92左右,平均召回率为0.913 1;F1值在0.89左右,平均F1值为0.888 7。因此LayoutLMv3-CSP实验结果中3项评价指标较未改进LayoutLMv3而言均有明显提升,且在一定范围内波动,也进一步说明了LayoutLMv3-CSP能够有效提升的相关评价指标。

表4 LayoutLMv3-CSP的K折交叉验证实验结果

Tab.4 K-fold cross verification of experimental results of improved LayoutLMv3

编号	精确率	召回率	F1值	编号	精确率	召回率	F1值	编号	精确率	召回率	F1值
1	0.872 3	0.896 8	0.884 4	3	0.865 6	0.919 8	0.891 9	5	0.853 3	0.908 6	0.880 1
2	0.865 3	0.921 1	0.892 9	4	0.870 2	0.919 1	0.894 0	平均值	0.868 8	0.919 0	0.893 1

3.4　LayoutLMv3-CSP 消融实验

预训练模型中不同预训练任务对模型的贡献程度可能会有所不同，本研究在 XFUND 数据集的基础上进行消融实验，从而研究每个预训练任务对模型整体的贡献程度，具体结果如表 5 所示。

表 5　不同预训练任务的实验结果

Tab. 5　Experimental results of different pre-training tasks

编号	预训练任务	精确率	召回率	F1 值	编号	预训练任务	精确率	召回率	F1 值
1	LayoutLMv3	0.847 0	0.919 1	0.881 6	5	CSP＋MIM＋MLM	0.849 2	0.927 9	0.886 8
2	CSP＋MIM	0.841 9	0.927 9	0.882 8	6	CSP＋MIM＋WPA	0.855 0	0.921 1	0.886 8
3	CSP＋MLM	0.857 2	0.934 5	0.894 2	7	CSP＋MLM＋WPA	0.848 2	0.924 0	0.884 5
4	CSP＋WPA	0.865 8	0.921 4	0.892 7	8	CSP＋MIM＋MLM＋WPA	0.875 1	0.924 7	0.899 2

为了研究模型中不同预训练任务的效果，本研究首先将 CSP 与 MIM、MLM 和 WPA 分别结合进行实验。与未改进 LayoutLMv3 相比，上述任务的结合使得语义信息抽取的各项指标均有所提升。尤其是 CSP 和 WPA 的结合使得预测语义信息抽取精确率、召回率和 F1 值分别提升了 1.88 百分点、0.23 百分点和 1.11 百分点，表明 CSP 和 WPA 在提升语义信息抽取的各项评价指标方面具有较大的贡献。

此外，为了研究模型中不同预训练任务组合对模型整体的贡献程度，本研究进行了 CSP 与其他 3 个预训练任务中的任意两者进行组合的实验。相对于未改进 LayoutLMv3，这些组合使得语义信息抽取的各项评价指标均有所提升。而将 CSP、MIM、MLM 和 WPA 全部进行组合后，预测的语义信息抽取精确率、召回率和 F1 值分别提升了 2.81 百分点、0.56 百分点和 1.76 百分点，表明 CSP、MIM、MLM 和 WPA 的组合在提升语义信息抽取的各项评价指标方面也具有一定的贡献。

4　结束语

本研究首先提出了一种基于单元格结构信息的表格识别算法，该算法充分利用单元格结构信息来解决端到端信息抽取不准确的问题，并验证了单元格结构信息在多模态预训练模型中的重要性。其次，本研究在 LayoutLMv3 中增加了新的预训练任务 CSP，优化了单元格结构信息在模型中的向量表示。最后，在最新的 LayoutLMv3 基础上进行了多任务扩展。实验结果表明，改进后的 2 种方法在表格语义信息抽取任务中取得了更好的效果。

在预训练模型中，不同预训练任务对模型的重要性可能会有所差异。将多任务学习视为多目标优化问题，并通过梯度下降求解到局部最优点，以进一步提高语义信息抽取的精确率，是未来的研究方向。

参考文献：

[1] ZHANG P，XU Y L，CHENG Z Z，et al. TRIE：end-to-end text reading and information extraction for document understanding [C]//Proceedings of the 28th ACM International Conference on Multimedia. Washington：ACM Digital Library，2020：1413-1422.

[2] YANG X，YUMER E，ASENTE P，et al. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu：IEEE，2017：4342-4351.

[3] SMITH R. An overview of the Tesseract OCR engine [C]//Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). Curitiba：IEEE，2007：629-633.

[4] YAO X C，van DURME B. Information extraction over structured data：question answering with freebase[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore：Association for Computational Linguistics，2014：956-966.

[5] HUANG Y P，LÜ T C，CUI L，et al. Layoutlmv3：Pre-training for document ai with unified text and image masking[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon：Association for Computing Machinery，2022：4083-4091.

[6] LI Y L，QIAN Y X，YU Y C，et al. Structext：structured text understanding with multi-modal transformers[C]//Proceedings of

the 29th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2021: 1912-1920.

[7] KATTI A R, REISSWIG C, GUDER C, et al. Chargrid: towards understanding 2D documents[EB/OL]. (2018-09-24)[2023-12-01]. https://arxiv.org/pdf/1809.08799.pdf.

[8] WANG H Y, CHENG Y H, CHEN C L P, et al. Semisupervised classification of hyperspectral image based on graph convolutional broad network[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 2995-3005.

[9] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2023-12-01]. https://arxiv.org/pdf/1810.04805.pdf

[10] XU Y H, LI M H, CUI L, et al. Layoutlm: pre-training of text and layout for document image understanding[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York: Association for Computing Machinery, 2020: 1192-1200.

[11] XU Y, XU Y H, LÜ T C, et al. Layoutlmv2: multi-modal pre-training for visually-rich document understanding[EB/OL]. (2020-12-29)[2023-12-01]. https://arxiv.org/abs/2012.14740.

[12] XU Y H, LÜ T C, CUI L, et al. Layoutxlm: multimodal pre-training for multilingual visually-rich document understanding[EB/OL]. (2021-09-09)[2023-12-01]. https://arxiv.org/abs/2104.08836.

[13] LI C L, BI B, YAN M, et al. Structurallm: structural pre-training for form understanding[EB/OL]. (2021-05-24)[2023-12-01]. https://arxiv.org/abs/2105.11210.

[14] OTTER D W, MEDINA J R, KALITA J K. A survey of the usages of deep learning for natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(2): 604-624.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, New York: Association for Computing Machinery, 2017: 6000-6010.

[16] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[EB/OL]. (2021-02-24)[2023-12-01]. https://arxiv.org/abs/2102.12092v1.

[17] SHAW P, USZKOREIT J, VASWANI A. Self-attention with relative position representations[EB/OL]. (2018-03-06)[2023-12-01]. https://arxiv.org/pdf/1803.02155.pdf.

[18] KIM W, SON B, KIM I. Vilt: vision-and-language transformer without convolution or region supervision[EB/OL]. (2021-05-05)[2023-12-01]. https://arxiv.org/abs/2102.03334v1.

Operations Research and Cybernetics

Table Extraction Method Incorporating Cell Structural Information

QIAO Yan[1], WU Zhiyou[1], GAO Huan[2], DUAN Xuxiang[3]

(1. School of Mathematical Sciences, Chongqing Normal University, Chongqing 401331;
2. Intel Joint Research Institute for Edge Intelligence, Nanjing 211135;
3. School of Mathematics and Statistics, Chongqing University, Chongqing 401331, China)

Abstract: In view of the fact that the existing end-to-end methods and pre-training model-based methods do not effectively utilize the structural information of the table cells during the training process, which affects the vector representation of the table text in the model and the final semantic information extraction accuracy, an end-to-end method that further utilizes the structural information of the cells to improve the effectiveness of the optical character recognition, and a pre-training method that increases the cell sequence prediction task are proposed. The experimental results show that the improved 2 methods achieve better results in the task of table semantic information extraction, with F1 values improved by 0.204 6 and 0.017 6. The improved methods reinforce the importance of cell structure information in tables and improve the accuracy rate of table semantic information extraction.

Keywords: table information extraction; cell structural information; table recognition algorithm; cell range recognition

(责任编辑 方 兴)