医疗材料光学字符识别要素提取数据集 MedOCR

刘利锋 常德杰 赵晓龙 王铁虎 杨锦新 郭龙杰 陈漠沙

（北京环球医疗救援 北京 100020） （阿里巴巴 杭州 310000）

汤步洲

（哈尔滨工业大学（深圳）鹏城实验室 深圳 518055）

〔摘要〕 介绍医疗材料光学字符识别要素提取数据集 MedOCR 设计目的、标注过程及数据特点，详细阐述 MedOCR 数据集的数据来源、标注方法、材料示例，分析数据集测评结果及应用情况，指出研究人员可基于 MedOCR 开展医疗材料信息提取方向的研究。

〔关键词〕 医疗行业；人工智能；信息提取

〔中图分类号〕 R－058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2022.12.006

MedOCR: The Dataset for Extraction of Optical Character Recognition Elements for Medical Materials *LIU Lifeng, CHANG Dejie, ZHAO Xiaolong, WANG Tiehu, YANG Jinxin, GUO Longjie, Beijing Universal Medical Assistance, Beijing 100020, China; CHEN Mosha, Alibaba Group, Hangzhou 310000, China; TANG Buzhou, Peng Cheng Laboratory, Harbin Institute of Technology, Shenzhen 518055, China*

〔**Abstract**〕 The paper introduces the design purpose, annotation process and data characteristics of MedOCR, an Optical Character Recognition (OCR) element extraction dataset for medical materials, elaborates the data sources, annotation methods, material examples of MedOCR dataset, and analyzes the evaluation results and application of the dataset. Researchers can carry out research in the direction of information extraction of medical materials based on MedOCR.

〔**Keywords**〕 medical industry; Artificial Intelligence (AI); information extraction

1 引言

1.1 MedOCR 设计目的和标注过程

1.1.1 设计目的 医疗行业、保险行业中的电子病历、照片等材料含有很多隐藏信息。基于人工智能技术合理利用这些信息，将在商业应用和科研领域产生很高的价值。这些信息（如客户信息、诊断信息、用药信息、费用信息等）需要进行结构化才能进一步使用。使用传统的文本识别方法，首先需要用计算机视觉（Computer Vision，CV）领域算法将图片材料进行文本化，其中包含目标检测和目标识别，如使用可微分二值化（Differentiable Binariza-

〔修回日期〕 2022－10－14

〔作者简介〕 刘利锋，首席执行官；通信作者：常德杰。

tion，DB）[1]算法进行目标检测，再使用卷积循环神经网络（Convolutional Recurrent Neural Network，CRNN）[2]进行目标识别，或者使用检测和识别结合的算法，例如点集网络（Point Gathering Network，PGnet）[3]进行文本化后，再使用自然语言处理（Natural Language Processing，NLP）领域的一些算法将这些文本信息结构化，例如使用双向编码器表征（Bidirectional Encoder Representations from Transformers，BERT）[4]进行文本分类，使用长短期记忆人工神经网络（Long - Short Term Memory，LSTM）[5]参与序列标注等一系列方法将文本结构化后根据需求使用。但是相较于传统机器学习，这些深度学习方法需要更多标注数据，如何获取高质量的标注数据成为相关研究进展的最大障碍，MedOCR 由此诞生，其目的在于提供高质量的标注数据以供材料信息提取相关研究使用。

1.1.2 标注过程 由业务专家设计标注指南，1 名主标注员和 1 名副标注员参考标注指南历时 1 个月标注完成。该数据集包含出院小结、购药发票、门诊发票、住院发票 4 类材料共1 700 张图片，涉及 87 个属性字段。

1.2 MedOCR 特点

1.2.1 数据特殊性 首先，相较于以往的光学字符识别（Optical Character Recognition，OCR）数据集（如 ICDAR 2017 - RCTW[6]采用四点标注），在 MedOCR 中，标注数据不涉及坐标，直接采用图片到值标注，拓展了研究方向，打破了传统文本识别的局限。其次，相较于以往的数据集，MedOCR 使用了医疗门诊发票、住院发票、购药发票、出院小结等 4 类病历材料。

1.2.2 数据复杂多样性 主要体现在两个方面，一是 OCR 场景复杂，包含打印模糊、打印偏斜、套打偏斜、套打覆盖、阴影覆盖等多种实际业务中会遇到的场景，对常规的 OCR 模型提出挑战；二是文本结构复杂，有上下结构文本、左右结构文本、无属性名、多种近似且非规范化属性名等复杂文本结构，这也是近年来结构化文本理解及命名实体识别（Named Entity Recognition，NER）的研究热点。

2 资料与方法

2.1 数据来源

MedOCR 的原始数据集数据源自互联网，将数据集分成训练集、评估 A 榜和评估 B 榜共 3 份，其中对应的材料类别数量，见表 1。

表 1 MedOCR 构成

数据集名称	分级	材料类别	类别数量
医疗标注任务数据集	训练集	出院小结	200
		购药发票	200
		门诊发票	200
		住院发票	400
	评估 A 榜	出院小结	40
		购药发票	40
		门诊发票	40
		住院发票	80
	评估 B 榜	出院小结	100
		购药发票	100
		门诊发票	100
		住院发票	200

2.2 标注方法

第 1 步：先从公开的互联网渠道找到 1 700 张 4 类病历材料，其中包括出院小结 340 张、购药发票 340 张、门诊发票 340 张、住院发票 680 张。第 2 步：选择特定字段，如购药发票特定提取字段：票据代码、票据号码、校验码、开票日期、收款人、复核人、价税合计（大写）、价税合计（小写）。第 3 步：进行严格的人工审核。第 4 步：得到图片对应的重要信息内容并以表格形式保存，见表 2、图 1。

表 2 购药发票提取字段格式

图名	材料类型	属性名	正确值
图片 1	购药发票	票据代码	值 1
图片 1	购药发票	票据号码	值 2
图片 1	购药发票	校验码	值 3
图片 1	购药发票	开票日期	值 4
图片 1	购药发票	收款人	值 5
图片 1	购药发票	复核人	值 6
图片 1	购药发票	价税合计（大写）	值 7
图片 1	购药发票	价税合计（小写）	值 8

图 1 数据集标注流程

2.3 各类材料数据示例说明

材料共有 4 类，分别是出院小结、购药发票、门诊发票和住院发票。每类材料标注数据文件结构基本一致，是由序号、图名、材料类型、属性名和正确值作为列名组成的 5 列数据，区别在于每类材料对应的属性名不同，出院小结共有 8 个属性名，购药发票共有 8 个属性名，门诊发票共有 34 个属性名，住院发票共有 37 个属性名，每个属性名对应 1 个正确值，即标注值，其根据图片上属性名对应的真实值标注而来。在正确值这列中有两个特殊值，分别为“无”和“-”，“无”代表图片中未出现该字段，“-”代表图片中出现该字段但没有对应值，见表 3。

表 3 部分标注数据

序号	图名	材料类型	属性名	正确值
1	F77ca7d0	出院小结	性别	-
2	F77ca7d0	出院小结	年龄	-
3	F77ca7d0	出院小结	医疗名称	哈尔滨医科大学附属第一医院
4	F77ca7d0	出院小结	组织机构代码	无
5	F77ca7d0	出院小结	医疗机构类型	无
6	F77ca7d0	出院小结	入院日期	2021-05-01
7	F77ca7d0	出院小结	出院日期	2021-05-02
8	F77ca7d0	出院小结	住院天数	5

3 结果评测

MedOCR 采用准确率作为评测指标。假如预测了一批字段，设预测正确的字段数量为 correct，预测错误的字段数量为 error，准确率就是预测正确字段数量占预测正确字段数量与预测错误字段数量之和的比例，当预测值和正确值皆为“无”时，则此值不计入计算，当预测值与正确值完全一致则判定为预测正确，否则判定为预测错误。具体计算方法见公式（1）。评估结果示例，见表 4，评估结果中 correct 为 3，error 为 4，则准确率为 0.42。

$$准确率 = \frac{correct}{correct + error} \quad (1)$$

表 4 评估结果

图名	材料类型	属性名	正确值	预测值	评估结果
图片 1	材料类型 1	字段 1	无	无	不计入计算
图片 1	材料类型 1	字段 2	无	11.00	预测错误
图片 1	材料类型 1	字段 3	-	-	预测正确
图片 1	材料类型 1	字段 4	-	女	预测错误
图片 2	材料类型 2	字段 1	100.00	100.00	预测正确
图片 2	材料类型 2	字段 2	20.00	21.00	预测错误
图片 2	材料类型 2	字段 3	长安医院	长谷医院	预测错误
图片 2	材料类型 2	字段 4	男	男	预测正确

4 数据集应用

4.1 应用方法及结果

正常未经过训练的模型对生活场景下病历材料图片内的文本识别效果普遍不佳，而且提取内容连贯性较差，丧失原本语义，因此需要在提取时考虑其原有内容完整性。评测组织者的基线方法（baseline）使用 DB[1] 进行文本检测和 CRNN[2] 进行文本识别的两阶段算法实现 OCR。为解决图片不规整造成的提取困难，使用 BERT 进行文本分类以及 LSTM + CRF 进行命名实体识别的方法处理 OCR 提取后的文本部分。利用上述优化后的模型对此数据集进行测试，见表 5。

表 5 模型测验结果

材料类型	准确率
出院小结	0. 884
购药发票	0. 861
门诊发票	0. 793
住院发票	0. 920
合计	0. 883

4.2 主要错误类型

经过对原数据图片和输出值的对比，主要错误类型可以总结为以下几类。第 1 类是印章对底版的信息遮挡导致目标提取有误，在各类发票中，各机构会盖上红色或蓝色公章，这些公章掩盖了部分底版信息，导致信息不能正确提取。第 2 类是底版背景颜色深对显示不清晰的信息造成影响，部分发票底版为深色背景，同样会导致信息提取产生误差。第 3 类是票据中常出现套打情况，即打印非一次完成而是票据内容打印在印好的票据底版上。套打时会因放置票据偏斜造成票据内容覆盖打印在底版文字上和打印内容偏离指定区域两种情况。第 4 类是信息内容不清晰，造成检测识别模型准确率低。第 5 类是检测模型误差造成识别模型识别不准确。第 6 类是识别模型混淆相近字造成识别结果不正确。第 7 类是文本结构不统一，且其中各省市票据字符、语义信息排列不规律造成 NER 模型训练提取难度提升。第 8 类是前置处理流程中检测识别模型结果不正确，造成错别字和文本位置信息错误，导致 BERT 结构化模型提取失败。从整体实验结果及错误类型分析中可以看到，目前模型性能相比人工标注结果还有较大的提升空间，信息抽取效果有待提高。

5 结语

本文介绍了专门用于医疗病历材料标注的特殊数据集。通过认真严格的标注过程获得了高质量数据集。实验结果证明医疗标注任务数据集具有一定实用性，但其信息抽取效果有待提高。该数据集的发布有助于推进医疗信息提取 OCR 模型的优化，并促进人工智能技术在医疗领域的应用。

参考文献

1 Liao M, Wan Z, Yao C, et al. Real - time Scene Text Detection with Differentiable Binarization [EB/OL]. [2022 - 06 - 20]. https: //arxiv. org/pdf/1911. 08947v2. pdf.

2 Shi B, Bai X, Yao C. An End - to - End Trainable Neural Network for Image - based Sequence Recognition and Its Application to Scene Text Recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (11): 2298 - 2304.

3 Wang P, Zhang C, Qi F, et al. PGNet: Real - time Arbitrarily - Shaped Text Spotting with Point Gathering Network [EB/OL]. [2022 - 06 - 20]. https: //arxiv. org/pdf/2104. 05458. pdf.

4 Devlin J, Chang M W, Lee K, et al. BERT: Pre - training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. [2022 - 06 - 20]. https: //arxiv. org/pdf/1810. 04805. pdf.

5 Hochreiter S, Schmidhuber J. Long Short - Term Memory [J]. Neural Computation, 1997, 9 (8): 1735 - 1780.

6 Shi B, Yao C, Liao M, et al. ICDAR 2017 Competition on Reading Chinese Text in the Wild (RCTW - 17) [EB/OL]. [2022 - 06 - 20]. https: //arxiv. org/pdf/1708. 09585v1. pdf.

(上接第 27 页)

8 刘苏文, 邵一帆, 钱龙华, 等. 基于联合学习的生物医学因果关系抽取 [J]. 中文信息学报, 2020, 34 (4): 60 - 68.

9 Shang Y M, Huang H, Mao X L. OneRel: Joint Entity and Relation Extraction with One Module in One Step [EB/OL]. [2022 - 06 - 30]. https: //arxiv. org/pdf/2203. 05412v1. pdf.

10 Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need [C]. Long Beach: The 31st International Conference on Neural Information Processing Systems, 2017.

11 Devlin J, Chang M W, Lee K, et al. BERT: Pre - training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. [2022 - 06 - 30]. https: //arxiv. org/pdf/1810. 04805v1. pdf.

12 Zheng H, Wen R, Chen X, et al. PRGC: Potential Relation and Global Correspondence Based Joint Relational Triple Extraction [EB/OL]. [2022 - 06 - 30]. https: //arxiv. org/pdf/2106. 09895v1. pdf.