中文医疗因果关系抽取数据集 CMedCausal

李子昊 陈漠沙 马镇新 尹康平 童毅轩 谭传奇 郎珍珍

（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）

徐 健

（阿里巴巴 杭州 310000）

〔摘要〕 介绍医学领域实体、关系抽取相关研究情况，详细阐述中文医疗因果关系抽取数据集 CMedCausal 构建方法及实验情况，提出利用数据集定义 3 类关键的医学因果推理关系：因果关系、条件关系和上下位关系，研究人员可基于 CMedCausal 开展医疗因果关系挖掘、医疗因果解释图谱构建等方向的研究。

〔关键词〕 因果关系；关系抽取；解释性；人工智能；自然语言处理

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.12.005

CMedCausal: Chinese Medical Causal Relationship Extraction Dataset *LI Zihao, CHEN Mosha, MA Zhenxin, YIN Kangping, TONG Yixuan, TAN Chuanqi, LANG Zhenzhen, Alibaba Group, Hangzhou 310000, China; TANG Buzhou, Peng Cheng Laboratory, Harbin Institute of Technology, Shenzhen 518055, China; XU Jian, Alibaba Group, Hangzhou 310000, China*

〔**Abstract**〕 The paper introduces relevant studies on entity and relation extraction in the medical field, elaborates the construction method and experimental results of the Chinese medical causal relationship extraction dataset—CMedCausal, and proposes to define three key types of medical causal explanation and reasoning relationships by using the dataset: causal relationship, conditional relationship and hypothetical relationship. Researchers can conduct medical causal relationship mining and medical causal interpretation map construction based on CMedCausal.

〔**Keywords**〕 causal relationship; relation extraction; interpretability; Artificial Intelligence (AI); Nature Language Processing (NLP)

1 引言

互联网在线问诊文本中包含大量医学相关概念，如何利用文本挖掘和深度学习技术获取相关医学知识近年来受到广泛关注[1－2]。然而医学概念的复杂性和多样性、医疗数据的隐私性都为相关研究带来巨大挑战。近年来，国际生物与临床信息学集成研究项目（Informatics for Integrating Biology and the Bedside，i2b2）以及中国健康信息处理会议（China Health Information Processing Conference，CHIP）等积极倡导从医疗数据中挖掘相关信息，针

〔修回日期〕 2022－10－31

〔作者简介〕 李子昊，硕士研究生；通信作者：陈漠沙，发表论文 20 篇。

对非结构化病历数据组织一系列评测任务，这些评测任务和数据集在相关研究社区中获得广泛影响力，在医学信息处理领域发挥了重要作用。

医学领域的实体、关系抽取技术可识别医学概念以及概念之间的相互关系，并将这些知识应用到医疗知识图谱中，从而能有效提升医疗图谱的可解释性。人工标注图谱成本较高，为了获取更多、更准确的关系知识，需要利用实体关系联合抽取技术[3-5]。

因果关系是一种重要的关系类型，特别是在注重可解释性的医学领域文本中。目前国外研究人员已提出多个因果关系抽取数据集，如 Dominique M 等[6]提出的基于金融领域的因果抽取数据集 FinCausal，Tan F A 等[7]提出的基于新闻领域的因果关系提取任务，在医疗领域 BioCreativeV 社区提出的从生物医学文献中自动抽取因果关系实体并用相关语句表示的任务[8]。相较于国外，国内医学因果关系推理方面的公开数据集资源还比较匮乏。因此，本文充分利用医学搜索引擎以及在线问诊的医疗回答文本，构建首个中文医学因果关系抽取数据集 CMedCausal，并依托 CHIP 2022 会议举办“医学因果实体关系抽取”评测比赛（http://cips-chip.org.cn/2022/eval2）。研究人员可利用 CMedCausal 开展医学因果关系挖掘，因果解释网络构建等方向的工作，从而提升医疗问诊结果的可解释性。

2 数据集构建

2.1 数据来源

抽取有来医生网站（https://m.youlai.cn）上较为工整且长度超过 200 个中文字符的线上问诊及医典百科数据。所采集大部分网上公开问诊数据并没有涉及患者隐私信息，所以不需要进行脱敏处理。筛选后的文本共包含 9 153 段文本，文本平均长度为 265 个字符。

2.2 任务定义

2.2.1 概述 数据集需要对医学概念片段以及医学概念片段之间的关系进行标注。医学概念片段指可作为一个独立语义单元的连续字符片段，可以是医学实体、临床发现或者具体疾病症状，从因果谓词表达上看这些片断行使条件、原因或者结果的语义角色，边界通常采用奥卡姆剃刀原则，保留原始含义的最小片段。标注人员限定了以临床发现和疾病为中心的医学概念片段内容，临床发现也包括实验室检验结果以及检查结果。医学概念片段之间关系包括因果关系、条件关系、上下位关系 3 种类型。

2.2.2 因果关系 指某种原因直接导致某种结果的关系。对于医学上常见的疾病和临床之间的关系即归类为因果关系。例如“人体的胃肠道功能紊乱，导致患者吸收能力变差”。本例中“胃肠道紊乱”是一个医学概念片段，“胃肠道功能紊乱”是“吸收能力变差”的直接原因，“吸收能力变差”是“胃肠功能紊乱”的直接结果。因果关系是医疗问诊里最常见的关系，也是判断问诊回答逻辑性最重要的依据，对于构建整个医疗知识图谱、实现自动诊断、提高医疗问诊可解释性有重要意义。

2.2.3 条件关系 指医学概念片段中一些特定的条件，用于修饰特定的因果关系。例如，“对阿莫西林过敏的患者不可以使用，服用阿莫西林可能会引起皮疹、药物热和哮喘等过敏反应，因此使用前一定要做青霉素皮试试验”。本例中“对阿莫西林过敏”是“服用阿莫西林”导致“皮疹”的条件。与因果关系不同的是，条件概念片段并不能直接导致某个结果发生。

2.2.4 上下位关系 指医学概念中的大小和蕴含关系，一般指某个宽泛、总称概念包含某个具体、特殊概念，例如，“阿尔茨海默症是一种精神类疾病”，本例中“精神类疾病”包含了“阿尔茨海默症”这一特定的精神类疾病。上下位关系是医学概念中较为重要的关系，对于医学概念的分类、医学图谱构建有重要作用。

2.3 数据标注

2.3.1 标注规范 准则 1：医学概念片段应尽可能包含完整有用的信息，包括症状的程度、频率

等，无关信息不在标注范围内。如“不及时治疗在局部可能会引起疼痛”中需标注“局部可能会引起疼痛”，仅标注“疼痛”则存在信息丢失；如果涉及人群信息来区分疾病特点，则需要标注人群，如“小儿咳嗽”。准则 2：针对多个医学概念片段组合在一起的长实体，采用如下约定进行标注。若每个概念片段具备独立意义则分开标注，如“过量饮酒、使用激素、劳累等引起的股骨头缺血性病变”中标注（“过量饮酒”，“股骨头缺血性病变”）（“使用激素”，“股骨头缺血性病变”）和（“劳累”，“股骨头缺血性病变”）3 对因果关系；若为非连续实体则合并标注，如“食用奶酪、巧克力、可乐会导致过度肥胖”中标注（“食用奶酪、巧克力、可乐”，“过度肥胖”）这对因果关系。其中非连续实体是指多个实体共用部分文字进而导致实体不连续的现象，例如上面例子中，“食用巧克力”和“食用可乐”即属此类。准则 3：任务只标注直接关系，不标注间接的推导关系，例如“A 导致 B，B 又导致 C”，则本任务只标注（A，B）和（B，C）两对因果关系，（A，C）不做标注；同样的，对于上下位或者别名的情况，仅标注最直观的实体，如“A，又称为 B，会导致 C”，只标注（A，C）。

2.3.2 标注过程 本任务由 1 名医学专家、1 名人工智能算法专家带领 8 名医学院本科生基于阿里巴巴夸克内部的标注平台完成，前后用时 1.5 个月。标注流程分为 4 个主要阶段，见图 1。（1）标注规范制定。规范主要由医学专家制定，在此阶段算法专家从模型处理能力的视角对规范提出优化建议，如医学专家倾向于将多个医学概念组合在一起标注为一个长实体片段，算法专家则会根据模型经验建议将其标注为独立意义的片段（参见标注规范准则 2）。最终目标是保证标注规范既符合医学常识，同时也对算法模型友好。（2）试标注。在试标注阶段医学专家会对 8 名医学院本科生进行系统性的任务讲解和规范培训，并带领 8 名医学生每人完成 20 条数据标注，目标是帮助标注人员充分理解任务，并能快速熟悉标注工具。接下来 8 名医学生和 1 名算法专家每人要独立完成 50 条数据的标注，在此期间医学专家会及时跟进标注人员遇到的问题，确保每位标注人员能充分理解任务并正确完成标注工作，同时也会根据标注人员的问题和反馈来优化标注规范。试标注阶段结束后，标注规范也最终定稿。该阶段耗时 1.5 周。（3）正式标注。由 8 名经过培训的医学生完成剩余语料标注，每人分配 1 080 条语料，8 位标注人员虚拟分为 4 个小组，同组内的两名标注同学之间有 100 条重复语料。这样设置的目的是为了统计和评估标注一致度。该阶段标注人员可以在标注工作组中提问和讨论问题，医学专家每天定时解答标注问题，并针对出现的共性问题组织讨论会。该阶段耗时 3 周。（4）质检。医学专家从每位标注同学的标注结果中随机挑选 50 条进行质检，分析标注错误类型并要求标注人员进行修复。质检阶段用时 1.5 周，经过 3 轮质检后（5 名标注人员经过两轮质检后验收合格，另 3 名经历 3 轮质检后验收合格），产出最终的 CMedCausal 数据集。

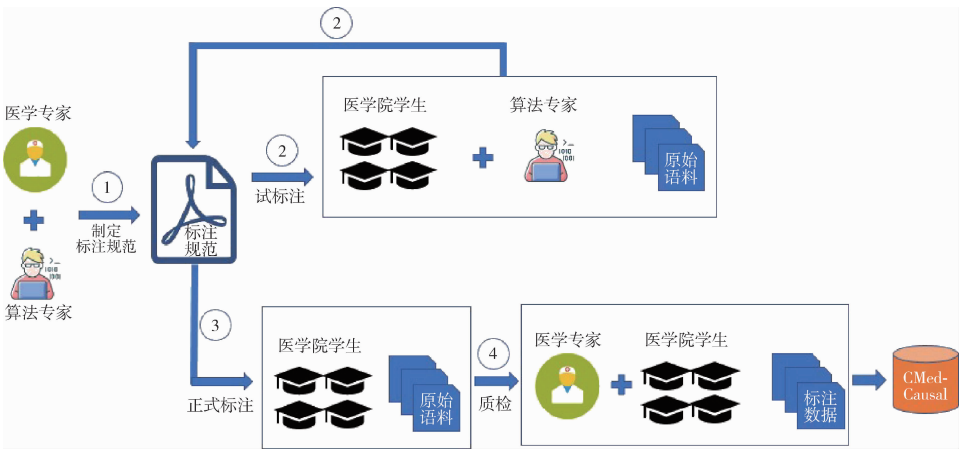

图 1 CMedCausal 标注过程

2.3.3 标注一致性 标注一致性（Inter - Annotator Agreement，IAA）是通过计算同一个虚拟小组内两名标注人员重叠标注的 100 条语料的微平均 *F*1 值（Micro - *F*1）和宏平均 *F*1 值（Macro - *F*1）指标来评估的。取 4 个虚拟小组的平均值得到的结果是：Micro - *F*1 是 0.741，Macro - *F*1 是 0.723。

2.3.4 数据统计 因果关系、条件关系和上下位关系 3 类关系的标注数量分别为 70 564、3 819 和 4 861，3 种关系占比分布为 18.5∶1∶1.3。

3 实验

3.1 实验数据

将实验数据按 8∶1∶1 的比例划分成训练、验证和测试集，并针对 3 份数据信息进行统计，见表 1。

表 1 实验数据信息统计

数据划分	样本数	字符总长度	平均字符长度	因果关系数	条件关系数	上下位关系数	关系比例
训练集	7 355	1 947 801	265	55 457	3 042	3 914	18.23∶1∶1.29
验证集	915	248 761	271	7 096	375	469	18.92∶1∶1.25
测试集	916	234 745	256	8 011	402	478	19.92∶1∶1.19

3.2 评价指标

本任务采用准确率（Precision，P）、召回率（Recall，R）和 *F*1 值（F - Measure，*F*1）作为评估指标。考虑到 3 类关系的比例相差较大，因此本任务采用 Macro - *F*1 作为最终评价标准。具体定义，假设有 n 个类别 C_1 、C_i 、C_n ，计算公式如下：设正确预测为类别 C_i 的样本个数为 $T_{p(i)}$，预测为 C_i 的样本个数为 T_i ，真实的 C_i 的样本个数为 P_i 。

$$Precision_i = \frac{T_{p(i)}}{T_i} \tag{1}$$

$$Recall_i = \frac{T_{p(i)}}{P_i} \tag{2}$$

$$Macro - F1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2 * Precisio\ n_i * Recall_i}{Precision_i + Recall_i} \tag{3}$$

3.3 实验环境及参数设计

本次实验选择两种常用的关系抽取（Subject - Predict - Object，SPO）模型作为基线（baseline）。OneRel：Shang Y M 等[9]提出的一种基于 Transformer[10]的 BERT[11]编码，后使用 N 个矩阵进行全局解码的模型，其中 N 为关系数，矩阵使用 3 种标记类型来记录 S 和 O 的起始终止位置关系。PRGC：Zheng H 等[12]提出的一种基于 BERT 编码，后判断文本的关系种类，根据关系种类预测 S 和 O，再使用对齐矩阵对相应的 S 和 O 进行对齐。编码器使用的是 bert - base - chinese（https：//huggingface.co/bert - base - chinese）模型，每批数据量设置为 6，迭代次数设置为 200，句子最大长度设置为 512。

3.4 实验结果分析

实体关系总体抽取结果，见表 2；3 类关系的具体实验结果，见表 3。

表 2 实验结果

模型	准确率	召回率	*F*1 值
OneRel	0.468	0.371	0.414
PRGC	0.411	0.207	0.275

表 3 3 类关系实验结果（OneRel/PRGC）

关系种类	准确率	召回率	*F*1 值
因果关系	0.50/0.41	0.35/0.22	0.41/0.28
条件关系	0.09/0.01	0.01/0.01	0.02/0.01
上下位关系	0.48/0.44	0.37/0.20	0.42/0.27

整体预测效果 OneRel 模型 *F*1 保持在 0.4 以上，PRGC 由于 *F*1 较低。随机抽取 100 条预测结果进行分析，发现错误类型可归结为 3 类。第 1 类错误是实体边界识别错误导致的，占比约 15%，如“血虚

型瘙痒症”可导致“皮肤可呈现大理石纹样”和“瘙痒剧烈”，由于这两个结果在文中是连在一起出现的，两种模型均将其识别为“血虚型瘙痒症”可导致“皮肤可呈现大理石纹样，瘙痒剧烈”。此类错误中模型会将两个或多个并列实体片段预测为一个长实体，导致召回率降低。第 2 类错误集中在特定修饰语的识别上，如“颈部淋巴结肿大”导致“脖子结节”，模型仅预测了“结节”这个结果，缺失了发病部位“脖子”，此类错误占比约为 20%。第 3 类错误主要分布在条件关系类别中，条件关系相比其他两类关系构成较为复杂，其尾实体是一个嵌套定义的因果关系，两种模型均无法很好地建模嵌套关系，如“前列腺增生”会导致“排尿不畅”，标注语料中“中老年男性”是该因果关系的修饰条件，但两种模型均只预测了（“前列腺增生”，“排尿不畅”）这对因果关系，无法准确捕获“中老年男性”这个修饰条件。从实验结果来看，条件关系最难预测，*F*1 分数不到 0.1。因此如何能同时正确识别出条件关系的头、尾实体是非常有挑战性的任务，进一步体现了 CMedCausal 数据集的难度。从整体实验结果及错误类型分析中可以看出，当前深度学习模型相比人工标注结果还有较大的提升空间，有待于探索更优的模型以及结合医学知识来达到更好的识别效果。

4 结语

医疗文本的因果实体关系抽取技术有助于提升医疗诊断整体逻辑性和可解释性，对于自动化问诊有重要作用，在此基础上可以进一步构建医疗知识图谱，从而挖掘更多的潜在关系。目前中文医疗因果关系抽取数据集较为缺乏，因此构建一个完善的关系抽取数据集对领域技术的发展有重要意义。

本文构建了一个专门用于医疗因果推断领域研究的因果关系抽取数据集 CMedCausal，系统地介绍了数据来源、标注规范及标注过程。数据集包含医学因果推断方面最常见的 3 类关系：因果关系、条件关系和上下位关系。CMedCausal 的构建方法具有一定有效性，为构建医疗知识图谱、医学因果解释网络、提高医疗问答可解释性奠定基础。

通过实验结果可以看出 CMedCausal 具有较高的挑战性，特别是条件关系的判断涉及复杂的因果关系推理以及修饰限定词的识别。同时相较于英文数据集，中文数据集由于词语之间无明显界限使得标注较为复杂，有无修饰语以及实体片段之间是否并列等情况均会导致预测结果与标注结果不一致，但并不能完全表明模型预测结果是完全错误的，因此如何建立一个合理的适用于医学因果关系抽取任务的评价标准也是至关重要的，有待于进一步探索。

参考文献

1 Raja U, Mitchell T, Day T, et al. Text Mining in Healthcare. Applications and Opportunities [J]. Journal of Healthcare Information Management, 2008, 22 (3): 52 – 56.

2 Esteva A, Robicquet A, Ramsundar B, et al. A Guide to Deep Learning in Healthcare [J]. Nature Medicine, 2019, 25 (1): 24 – 29.

3 Uzuner Ö, South B R, Shen S, et al. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text [J]. Journal of the American Medical Informatics Association, 2011, 18 (5): 552 – 556.

4 咎红英, 关同峰, 张坤丽, 等. 面向医学文本的实体关系抽取研究综述 [J]. 郑州大学学报 (理学版), 2020, 52 (4): 1 – 15.

5 Chang D, Chen M, Liu C, et al. DiaKG: An Annotated Diabetes Dataset for Medical Knowledge Graph Construction [EB/OL]. [2022 – 06 – 30]. https://arxiv.org/pdf/2105.15033.pdf.

6 Dominique M, Labidurie E, Ozturk Y, et al. Data Processing and Annotation Schemes for FinCausal Shared Task [EB/OL]. [2022 – 06 – 30]. https://arxiv.org/pdf/2012.02498v1.pdf.

7 Tan F A, Hürriyetolu A, Caselli T, et al. The Causal News Corpus: Annotating Causal Relations in Event Sentences from News [EB/OL]. [2022 – 06 – 30]. https://arxiv.org/pdf/2204.11714v1.pdf.

（下转第 31 页）

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.