基于中文病例报告文献的医学诊疗命名实体识别研究*

夏光辉 李军莲 邢宝坤 崔胜男

（中国医学科学院医学信息研究所 北京 l00020） （中国医学科学院北京协和医院 北京 100730）

〔摘要〕 基于公开发表的中文病例报告文献构建医学诊疗实体语料库，搭建语料标注审核平台，以基于上下文语义理解的方式识别疾病、症状、检查、治疗 4 类医学诊疗实体。通过构建字符、词边界、上下文、词性和词典等特征，基于条件随机场模型提出一种多特征融合的中文病例报告诊疗命名实体识别方法，具有较高的识别准确率。

〔关键词〕 中文病例报告；医学诊疗；命名实体识别；条件随机场

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2019.06.012

Study of Named Entity Recognition in Medical Treatment Based on Literatures of Chinese Case Reports *XIA Guanghui, LI Junlian, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijng 100020, China; XING Baokun, CUI Shengnan, Peking Union Medical College Hospital, Chinese Academy of Medical Science & Peking Union Medical College, Beijing 100730, China*

〔**Abstract**〕 Published literatures of Chinese case reports is used to build the medical treatment entity corpus and the auditing platform of corpus tagging to identify 4 medical treatment entities of diseases, symptoms, examinations and treatments in line with the semantic understanding of the context. A multi－feature named entity recognition approach for Chinese case reports is put forward based on the conditional random fields model through establishing features of characters, word boundary, contexts, and dictionaries. This approach is of higher recognition accuracy.

〔**Keywords**〕 Chinese case report; medical treatment; Named Entity Recognition (NER); Conditional Random Field (CRF)

1 引言

1.1 研究背景与必要性

2011 年 Watson 在美国最受欢迎的智力问答电视节目中亮相，一举打败人类智力竞赛的冠军，使人们认识到医疗人工智能应用前景广阔，智能医疗逐渐成为计算机和医学领域共同的研究热点。基于人工智能的医疗服务系统可以有效缓解优质医疗资源缺乏、分布不均而导致的看病难、医患关系紧张等问题，而建设智能医疗服务系统需要将海量的医疗数据转变为计算机可识别和计算的结构化形式，如何使计算机理解医疗大数据文本中的自然语言已

〔修回日期〕 2019－05－14

〔作者简介〕 夏光辉，硕士，助理研究员，发表论文30 余篇。

〔基金项目〕 国家科技图书文献中心“下一代国家科技创新知识服务开放系统”先期研发任务课题“文本知识对象语义标注研究”（项目编号：XQYF0201）；中国医学科学院医学与健康科技创新工程重大协同创新项目“生物医学科技信息支撑平台”（项目编号：2016－I2M－2－005）。

经成为智能医疗领域信息处理和数据挖掘研究面临的关键问题。

1.2 命名实体识别相关研究

命名实体识别（Named Entity Recognition, NER）是指从非结构化文本中抽取表达特定含义的实体，从而形成结构化、有明确类别归属的实体数据。中文医学领域的 NER 研究主要针对生物医学文献和电子病历两类文本，近年来基于中文电子病历的实体识别研究已经有比较多的成果，基于条件随机场（Conditional Random Field, CRF）的实体识别方法成为主流。目前国内研究者基于 CRF 模型提出多种中文电子病历实体识别的优化改进方案。许源等[1]针对脑卒中专科的 500 份入院记录构建语料库，采取基于 CRF 和规则相结合的方法进行医学实体识别。孙安等[2]以 CCKS2017 提供的 400 份电子病历数据作为研究对象，使用 CRF + + 工具，通过构建字粒度词语特征提升实体识别模型的性能。于楠等[3]以 400 份电子病历构建语料库，采取 CRF + + 工具进行实体识别，增加引导词特征和构词结构特征提升识别的准确性。张祥伟等[4]以 100 份中文电子病历构建语料库，基于 CRF 模型，构建语言符号、词性、关键词、词典、词聚类等多种特征识别疾病、症状、检查和治疗 4 类实体。杨红梅等[5]以 240 份肝细胞癌患者入院记录和出院小结构建语料库，采取长短期记忆网络（Long Short Term Memory, LSTM）与 CRF 相结合的方法构建命名实体识别模型。此外国内学者基于条件随机场模型的中文电子病历实体识别研究还包括电子病历中时间类实体信息抽取研究[6-7]、充分利用未标注语料的半监督学习方法研究[8]等。综上研究，条件随机场不仅可以使用字、词、词性等多种上下文特征，还可以灵活引入词典等外部特征，在命名实体识别任务中的效果已被广泛认可。但是由于患者隐私问题，电子病历难以获取，国内还没有公开可获得的电子病历数据库，因此研究者只能局限于在较小规模的数据集上进行算法验证，尚不能有效验证条件随机场模型在医学全学科进行命名实体识别的泛化能力。

1.3 病例报告

病例报告是医学论文的一种常见体裁，往往通过对 1 个、2 个或系列病例的诊疗经过进行生动记录和描述，试图在疾病的表现、机理以及诊断治疗等方面提供第一手感性资料[9]。病例报告类论文一般关注于一些首次发现或罕见、治疗相关的副作用以及多种症状重叠容易误诊的病例，病例报告中的病例资料是经过遴选、编辑、审校后的高质量病历数据。针对这些公开的优质病例资料自动识别与提取相关诊疗信息，不仅能为公众提供精准的健康信息服务，还能为临床决策支持、辅助问诊等应用场景提供数据支持。本文以公开发表的病例报告文献中的临床资料构建医疗实体识别语料库，使用条件随机场模型，融合多种特征，实现疾病、症状、检查、治疗等医疗实体的识别。

2 材料与方法

2.1 语料数据来源

本研究使用的病例报告原始语料均来源于中华医学会系列期刊 2015 年发表的 300 篇相关文献，为尽量保证病例资料覆盖医学全领域，分别从 11 种期刊中随机选择文献，获取病例报告文献的标题和临床资料两部分内容，具体情况，见表 1。

表 1 病例报告语料数据分布情况

刊名	文献量（篇）
《中华神经科杂志》	39
《中华放射学杂志》	48
《中华口腔医学杂志》	13
《中华内分泌代谢杂志》	15
《中华内科杂志》	22
《中华儿科杂志》	22
《中华外科杂志》	17
《中华妇产科杂志》	10
《中华眼科杂志》	41
《中华皮肤科杂志》	23
《中华医学杂志》	50
合计	300

2.2 命名实体分类设计

I2B2 2010 语料的实体类型分为医疗问题

（medical problem）、检查（test）和治疗（treatment）3 类[10]；杨锦锋等构建的电子病历语料库中实体类型分为疾病、疾病诊断分类、症状、检查、治疗 5 类[11]。本研究借鉴以上语料库构建经验，将病例报告中的实体分为疾病、症状、检查和治疗 4 种类型，参考一体化医学语言系统（Unified Medical Language System，UMLS）的语义类型界定每一类实体涵盖的范围，但不局限于 UMLS 中的概念。疾病（diseases）是导致患者处于非健康状态的原因或者医生对患者做出的诊断统称为疾病，如高血压、股骨骨折、畸形等。症状（symptoms）是指患者主观感受到的不适应或痛苦的异常感觉或某些客观的病态改变，同时还包括医师或其他人客观检查到的改变，即体征（sign），如肢体无力、耳鸣、恶心、低血压等[12]。检查（test）指的是为证实患者是否具有某种疾病或者出现某种症状而进行体格、实验室、器械检查等过程以及相应的检查设备、项目，如尿常规、心电图、心肺听诊等。治疗（treatment）指的是为解决疾病或者缓解症状而施加给患者的治疗程序、干预措施、给予药品、手术操作，如输血、胰岛素、肺切除术等。本研究所定义的命名实体遵循 3 条基本原则：实体是意义完整的最小片段；实体间不重叠、不嵌套、不含有除顿号以外的标点符号；“顿号”、“伴”、“及”、“并”等表达并列关系的字符，基于上下文语境理解不可或缺时可作为实体的组成部分。

2.3 实体语料标注

考虑到病例报告涉及的内容专业性较强，语料标注采取规范制定团队预先形成语料标注规范指导标注人员人工标注，遇到疑惑时标注人员可将实体标注状态修改为存疑，待与规范制定团队讨论达成一致后，修改实体标注状态并完善标注规范。标注团队主要包括 1 名医院病案科研究人员、1 名医院医护人员，他们在工作中参与电子病历的书写和核查，具备足够的医疗知识，积累丰富的临床经验；规范制定团队包括两名自然语言处理领域相关研究人员。整个语料标注过程分为 4 轮，其中前两轮是预标注，第 3 轮是正式标注，第 4 轮是标注审核。预标注旨在培训标注人员，熟悉标注规范的同时理解语料标注的目的，逐步完善标注规范，解决标注人员的疑问。经过两轮预标注，两名标注者的一致性达到 90%，开始正式标注[13]。预标注共包含 50 份病例报告，每轮的 25 份病例报告由两名标注者分别独立标注，简称为 A 和 B。病例报告语料标注流程，见图 1。两名标注者通过标注平台，参照标注规范，单独完成 25 份病例报告的标注。通过平台的比较功能，标注人员与规范制定人员共同讨论，针对不一致的标注实体达成一致后可在平台上进行修改、确认和删除，完善标注规范并补充标注样例，指导后续标注工作。病例报告标注平台语料审核示例，见图 2。两名标注人员标注结果不一致，标注平台将以不同颜色的文字表示实体类别，同时通过添加底纹的方式突出显示不一致的实体文字，以便于讨论修改形成统一共识的标注规范和语料库。第 3 轮正式标注共包含 300 份病例报告文本，其中包括预标注的 50 份病例报告。正式标注由两名标注者共同完成。为确保人工标注的进度和质量，规范制定者可在平台上实时查看标注进度并对已完成的标注文本进行审核。

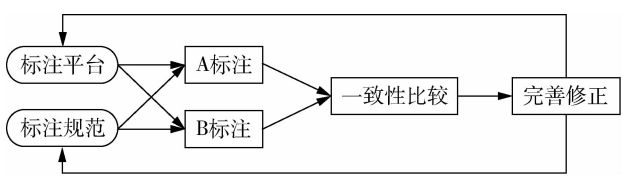

图 1 病例报告语料标注流程

图 2 病例报告标注平台语料审核

2.4 特征提取方案

实体特征是命名实体识别准确与否的决定性因素。在命名实体识别任务中常构建的实体特征包括字符、词性特征等，特征之间可以构成不同的组合。实体识别过程中关键在于针对相应任务模型选取合适、准确度高的特征来表示文本中隐含、嵌套的语言逻辑。医学病例报告文本一般采用叙述形式，具有语言简洁以及非标准的描述特性，因此没有高层次的句法特征。通过分析病例报告文本结构，本文从常用的特征集合中选取几种合适的来标识病例报告文本。（1）字符特征。是最基本、最直接地表达文本序列中元素的一类特征，本文所指的字符包括汉字、标点符号、外文字母、数字和日期等。（2）词边界特征。用于反映边界特征字符的位置信息，帮助确定命名实体的边界。本文实验中采用 BIEO 编码模式来表示输入观测序列元素的词边界特征。其中 B 表示实体名称的开始，即左边界；I 表示实体名称的内部，即实体的非边界部分；E 表示实体的结束，即右边界；O 表示非实体。（3）上下文特征。在本文中指的是窗口长度内观测值之间或特征之间的相互依赖关系，既可以表示实体内部的依赖关系，也可以表示实体内部与外部的相互关系。上下文窗口由当前词以及前后若干个词组成，上下文窗口长度依据所识别实体的长度进行设定[14]。在本文试验中窗口大小设置为 9。（4）词性特征。在自然语言处理任务中词性标记可以深度挖掘词语组合形成的词法信息，表达句子中存在的固有结构，提升特征集的区分度。词性特征由 ANSJ 中文分词工具生成。（5）词典特征。本文基于 CMeSH，依据主题词的树状结构号构造疾病、症状、检查、治疗等实体词典，基于实体词典采取字符串匹配的方式构建语料词典特征。（6）融合特征。当单一的特征不足以准确表达输入观测序列中元素之间的相互依赖关系时，可通过对不同单一特征的相应组合表示观测元素间更为深层次的语义结构。通过特征模板提出一种融合特征，由字符、词性和词典特征随机融合而成。

2.5 条件随机场模型

2001 年由 John Lafferty 等人提出基于统计的序列标注识别模型[15]。是连续优化的最大熵模型，具有较强的特征融合能力，可以在模型中灵活添加特征来表示元素之间的关系。克服观察值之间的独立假设，采用全局归一化的方法，有效避免数据稀疏性问题，特征权值全局最优，避免标注偏倚问题，在 CCKS 2017 中文电子病历命名实体识别评测会议中被广泛使用[16-17]。在条件随机场模型中，若令 $x = \{x_1, x_2, \cdots, x_n\}$ 为观测序列，$y = \{y_1, y_2, \cdots, y_n\}$ 为与之相应的标记序列，则条件概率为：

$$P(y \mid x) = \frac{1}{Z_x}\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(y_{i-1}, y_i, x, i)\right)$$

公式（1）

其中 Z_x 是所有状态序列的标准化因子，f_j 是特征向量函数，λ_j 是特征权重。当训练状态序列被完全明确地标记后可为该模型找到最优的 λ 值。在此基础上使用 Viterbi 算法得到最佳状态序列。本文使用斯坦福大学开源的命名实体类识别工具 Stanford

NER，其基于 CRF 模型实现。

3 试验结果与分析

3.1 评测结果

本文将 300 份病例报告语料中的 200 份作为训练集，100 份作为测试集。评价指标采用机器学习领域常用的准确率 P（Precision）、召回率 R（Recall）和 F 值（F－measure），具体计算公式如下：

$$P = \frac{TP}{TP + FP} = \frac{\text{正确识别的命名实体数}}{\text{识别出的命名实体数}} * 100\% \quad \text{公式（2）}$$

$$R = \frac{TP}{TP + FN} = \frac{\text{正确识别的命名实体数}}{\text{文本包含的命名实体数}} * 100\% \quad \text{公式（3）}$$

$$F = \frac{2 * P * R}{P + R} * 100\% \quad \text{公式（4）}$$

为验证不同特征对识别效果的影响，首先选择字符特征（word），然后逐渐增加词性特征（pos）、词典特征（dic）。不同特征的实验情况，见表 2。可以看出当添加词典特征时，相比于只用到字符特征，准确率略有上升，最终达到 79.40%，F 值也有提升，最终达到 76.67%。而添加词性特征时，相比于只用到字符特征，召回率略有上升，最终达到 74.30%，但准确率下降较多，最终 F 值为 76.44%。实验结果表明当融合字符、词典特征时实体识别的准确率能有效提高。而因病例报告中临床资料文本的语言特性，如行文中句子语法成分不完整、动词缺失，导致词性特征用于病例报告临床医学诊疗实体的识别效果不明显。

3.2 识别错误的实体

（1）在语料标注规范中基于临床含义和上下文语境的考虑，对语料中的顿号，及、和等表示并列关系的符号整体标注为诊疗实体，但基于语义理解的实体识别还有待进一步提高。如“计算力、记忆力差”识别为“记忆力差”，“脑脊液常规和生化”识别为“脑脊液常规”、“生化”两个检查实体，“双肾上腺 CT 平扫及强化”识别为“双肾上腺 CT 平扫”、“强化”两个检查实体。（2）语料数据稀疏导致实体未识别。如“神经功能”、“瞌睡”等实体未识别。（3）实体边界识别错误。如“双耳中低频感音性耳聋”识别为“感音性耳聋”，“阑尾炎术后”识别为“阑尾炎”。（4）实体类别识别错误。如误将治疗实体“膀胱肌瘤、大网膜肌瘤、肠系膜肌瘤、阑尾系膜肌瘤切除术”识别为疾病实体“膀胱肌瘤”、“大网膜肌瘤”、“肠系膜肌瘤”和治疗实体“阑尾系膜肌瘤切除术”。从以上分析得出，基于临床真实含义结合上下文语境进行命名实体识别、语料构建时标注人员也基于临床上下文语境标注诊疗实体，但本文构建的模型在进行基于语义理解的命名实体识别时精准度还有待提高。

表 2 采取不同特征组合的实体识别评测结果

特征	指标	疾病	症状	检查	治疗	总体
word	P	0.795 3	0.719 0	0.858 7	0.756 0	0.792 5
	R	0.699 8	0.697 3	0.849 1	0.575 3	0.741 2
	F1	0.744 5	0.708 0	0.853 9	0.653 4	0.766 0
word + pos	P	0.804 5	0.709 3	0.854 4	0.748 5	0.787 2
	R	0.694 5	0.698 1	0.847 3	0.595 2	0.743 0
	F1	0.745 4	0.703 7	0.850 8	0.663 1	0.764 4
word + dic	P	0.807 1	0.7205	0.856 9	0.758 8	0.794 0
	R	0.707 7	0.696 4	0.846 3	0.577 9	0.741 2
	F1	0.754 1	0.708 2	0.851 6	0.656 1	0.766 7
word + pos + dic	P	0.808 9	0.718 8	0.857 2	0.757 4	0.793 3
	R	0.700 2	0.700 1	0.842 3	0.577 6	0.739 8
	F1	0.750 6	0.709 3	0.849 7	0.655 4	0.765 6

4 结语

本文基于病例报告文献构建语料集，基于条件随机场模型提出一种多特征融合的中文病例报告诊疗命名实体识别方法。采用递增式的特征学习策略，测试不同特征组合用于中文病例报告命名实体识别的效果，最终融合字符、词边界、上下文、词性和词典等特征，构建的模型能准确识别出中文病例报告中的大部分诊疗实体。但由于病例报告中的病例资料高度凝练，其行文中并列结构、缩略语和上下文语义关联较多，该模型在此方面的识别能力还有待提高，后续可在以下几方面进一步优化：（1）将机器学习方法与基于语言规则的模式匹配方法相结合，提升针对特定语言结构的实体识别能力。（2）借助于本研究构建的病例报告标注规范和标注平台，构建更大规模的高质量医学诊疗语料库，提高机器学习模型的识别效果。（3）进一步优化机器学习算法，丰富中文语料特征集，结合医生临床诊疗过程的应用场景，研究基于语义理解的命名实体识别，推动实体识别技术在临床辅助诊疗决策中的应用。

参考文献

1 许源，葛艳秋，王强，等. 基于 CRF 与 RUTA 规则相结合的卒中入院记录医学实体识别及应用［J］. 中山大学学报（医学科学版），2018，39（3）：455－462.

2 孙安，于英香，罗永刚，等. 序列标注模型中的字粒度特征提取方案研究—以 CCKS2017：Task2 临床病历命名实体识别任务为例［J］. 图书情报工作，2018，62（11）：103－111.

3 于楠，王普，翁壮，等. 基于多特征融合的中文电子病历命名实体识别［J］. 北京生物医学工程，2018，37（3）：279－284，324.

4 张祥伟，李智. 基于多特征融合的中文电子病历命名实体识别［J］. 软件导刊，2017，16（2）：128－131.

5 杨红梅，李琳，杨日东，等. 基于双向 LSTM 神经网络电子病历命名实体的识别模型［J］. 中国组织工程研究，2018，22（20）：3237－3242.

6 孙健，高大启，刘珉，等. 中文电子病历文本中的时间识别算法研究［J］. 山西大学学报（自然科学版），2018，41（1）：15－22.

7 孙健，高大启，阮彤，等. 中文电子病历中的时间关系识别［J］. 计算机应用，2018，38（3）：626－632.

8 王润奇，关毅. 基于 Tri－Training 算法的中文电子病历实体识别研究［J］. 智能计算机与应用，2017，7（6）：132－134，138.

9 黄文华. 病例报告写作规范［J］. 神经病学与神经康复学杂志，2016，12（4）：228－232.

10 i2b2. Concept Annotation Guidelines［EB/OL］.［2019－01－18］. https：//www. i2b2. org/NLP/Relations/assets/Concept%20Annotation%20Guideline. pdf.

11 杨锦锋，关毅，何彬，等. 中文电子病历命名实体和实体关系语料库构建［J］. 软件学报，2016，27（11）：2725－2746.

12 （美）Marsha L Conroy. 临床症状与体征诊断指南［M］. 北京：科学出版社，201.

13 曲春燕，关毅，杨锦锋，等. 中文电子病历命名实体标注语料库构建［J］. 高技术通讯，2015，25（2）：143－150.

14 黄珺. 基于条件随机场的命名实体识别［D］. 桂林：桂林电子科技大学，2015.

15 Lafferty J，Mccallum A，Pereira F. Conditional Random Fields Probabilistic Models for Segmenting and Labeling Sequence Data［C］. WA，USA：Proceedings of the International Conference on Machine Learning，2001.

16 Jianglu H，Xue S，Zengjian L，et al. HITSZ_ CNER：a hybrid system for entity recognition from chinese clinical text［C］. Chengdu：Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing，2017：25－30.

17 Geng D W. Clinical Name Entity Recognition Using Conditional Random Field with Augmented Features［C］. Chengdu：Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing，2017：61－68.