信息系统学报
第 19 辑：62-71

China Journal of Information Systems
62-71

中文在线健康社区中的医疗命名实体识别方法研究*

杨杭州[1]　刘　凯[2]　颜志军[1]　李军莲[3]　孙海霞[3]
（1. 北京理工大学 管理与经济学院，北京　100081）
（2. 北京理工大学 信息化办公室，北京　100081）
（3. 中国医学科学院 医学信息研究所，北京　100020）

摘　要　医疗命名实体识别在网络非结构化健康文本的知识发现中至关重要。目前的方法存在难以识别网络上不规范的名称和需要人工标注等问题，且往往没有考虑医疗领域的特点。结合中文一体化医学语言系统和网络拓展词典构建的医疗领域词典，并使用基于语义规则的方法，本文提出一种基于 LDA 和 CRF 的混合模型识别在线医疗命名实体。与两种基准方法的比较表明了该方法在网络健康社区文本中医疗命名实体识别的有效性。

关键词　医疗命名实体识别，在线健康社区，LDA，CRF

中图分类号　C931.6

随着网络 2.0 的快速发展，越来越多的人参与到网络健康信息共享中来。其中，在线健康社区（online health communities，OHCs）是网络上人们寻找和分享健康信息的重要场所[1]。在 OHCs 中，病人可以向专业医生描述自己的病情并寻求医疗决策帮助，或寻找与自己症状或病情相似的求助者。医生可以根据病人的描述提供医疗建议或初步诊断。网络上病人和医生之间频繁的交互产生了大量的文本数据，其中包含了很多有价值的健康知识。这些知识对于辅助医疗决策、病人自我管理和药物安全警戒等有重要意义。但是在线医疗文本往往是大量的非结构化的数据，难以高效地从中提取出有用的信息。同时，人们关注的信息也有所不同，如病人较为关注与自己病情相关的内容，医生则更加关注自己所擅长的领域。

信息提取技术是解决上述问题的有效方法。信息提取的过程一般分为两个步骤：命名实体识别和实体关系提取。作为信息提取的基础，命名实体识别是自然语言处理领域中的一个重要研究内容。在医疗健康领域，医疗命名实体的识别对象主要包括以下三类：医疗问题、检查和治疗。例如，帖子“患有二型糖尿病，容易疲劳，以前服用过二甲双胍，现在空腹血糖是 7.5”。其中，“二型糖尿病”和“疲劳”是医疗问题类命名实体，“二甲双胍”是治疗类命名实体，“空腹血糖”是检查类命名实体。医疗命名实体识别是医疗领域自然语言处理任务中的重要步骤，许多医疗领域知识挖掘任务如药物不良反应的识别等都需要先提取出文本中的医疗命名实体，如基于网络文本的药物不良反应的发现任务，往往需要先提取出文本中的药物（治疗）类实体和不良反应（医疗问题）类实体，再进行实体关系的提取。因此，医疗命名实体识别的效果对医疗领域信息检索、信息提取和知识发现等任务的表现有至关重要的影响[2]。然而，目前的医疗命名实体识别方法还存在很多问题。首先，医疗相关的文本

* 基金项目：国家自然科学基金（71572013，71272057）。
通信作者：颜志军，北京理工大学管理与经济学院，教授，博士生导师，E-mail：yanzhijun@bit.edu.cn。

存在一些特殊的领域术语，而现有的很多命名实体识别方法并不是针对医疗健康领域，而是提取如人名、地名和组织名等实体类型。其次，现有从电子健康记录[3, 4]或者社交媒体[5, 6]中提取命名实体的研究并没有考虑网络环境的特殊性。网络环境一个明显的特点是文本内容的多样性。在电子健康记录中，医生一般会写下三种关于病人的信息，即医疗问题、检查和治疗。而在网络环境下可能会产生其他类型的信息，如关于饮食的信息，这些信息对医疗决策也会有所影响。

针对上述问题并结合在线医疗文本的特点，本文提出一种新的在线医疗命名实体识别的方法。基于中文一体化医学语言系统（Chinese unified medical language system，CUMLS）和网络词典构建的医疗领域词典，使用 NLPIR（natural language processing and information retrieval sharing platform，自然语言处理与信息检索共享平台，http：//ictclas.nlpir.org/）汉语分词系统作为分词工具，并结合基于语义规则的方法和 LDA 的主题模型，使用 CRF 方法建立医疗命名实体识别模型。本文的主要贡献在以下几个方面：考虑到在线健康社区中医疗文本的特殊性，使用网络词典和基于语义规则的方法识别网络中不规则的命名实体。同时，使用 LDA 和基于规则的方法能够实现文本的自动标注，避免了人工标注的过程。LDA 和 CRF 的特性使该方法能够适用于大规模的在线医疗数据处理任务。

1 文献综述

1.1 基于词典和规则的方法

目前命名实体的识别方法主要分为基于词典或规则的方法和机器学习的方法。基于词典的方法通过字符串的匹配实现命名实体的识别，但是对词典有较大的依赖性。目前国际上常用的词典主要包括国际疾病分类（international classification of diseases，ICD）、医学一体化语言系统（unified medical language system，UMLS）、医学系统命名法–临床术语（systematized nomenclature of medicine-clinical terms，SNOMED CT）和医学主题词表（medical subject headings，MeSH）等。Coden 等[7]提出一种可拓展和可修改的癌症疾病知识表示模型用于从病理报告中自动提取概念实体，该模型使用 ICD 作为基础词典，并基于规则的方法进行医疗术语的扩充和概念识别。Song 等[8]提出了一种基于 MeSH、UMLS 词典和 GENIA（一种基因语料）的生物医学实体提取技术。实验结果表明词典的选择对基于词典的命名实体识别方法有很大的影响。为了从社交媒体中提取药物的不良反应，Liu 等[6]使用了基于词典的方法提取药物和不良反应实体。除了 UMLS 和药物不良反应报告系统（food and drug administration，adverse event reporting system，FAERS），该方法还使用用户健康词表（consumer health vocabulary，CHV）来识别社交媒体中的不规范文本拼写。而对于中文的医疗命名实体识别，可使用的词典资源相对较少，这也在一定程度上制约了中文医学信息提取的研究。中文版本的医学主题词表（CMeSH）由中国医学科学院医学信息研究所翻译，但没有免费开放。我国卫生部也于 2012 年发布了中文版的 ICD-10，但并不适用于全面地提取各类医疗命名实体。另外，网络环境中文本内容的多样性也给基于词典和规则的方法带来了挑战。

1.2 机器学习的方法

因较好的环境适应性，基于机器学习的方法近年来被广泛应用在命名实体识别的任务中。在命名实体识别任务中，机器学习的方法大体上可以分为基于分类的方法和基于序列标注的方法。较为常用的分类方法包括支持向量机（support vector machine，SVM）、最大熵模型（maximum entropy model，ME）和结构化支持向量机（structural support vector machine，SSVM）等。SVM 是一种较为高效和流

行的分类方法，被广泛应用在命名实体识别及相关任务中。Jiang 等[9]将 SVM 和启发式算法结合，提出了一种医疗实体识别的混合方法。该模型考虑的特征包括词级别信息、句法信息、词典与语义信息和论述信息。结果表明词典与语义信息能有效提高该模型的识别效果。Saha 等[10]基于最大熵模型提出了生物医学命名实体识别的特征筛选方法，使用的特征包括特殊字符和词性信息等。

序列标注方法在命名实体识别的任务中被广泛应用，包括隐马尔可夫模型（hidden Markov model，HMM）、最大熵马尔可夫模型（maximum entropy Markov model，MEMM）以及 CRF。Sun 等[11]使用 CRF 从生物医学文献中提取医疗实体，考虑了正确拼写特征、上下文特征、词形特征、前后缀特征、词性特征和浅层句法特征。Lei 等[12]探究了不同类型的特征（包括分词、词性、分区信息等）和不同的机器学习的方法（包括 CRF、SVM、ME 和 SSVM）在中文医疗命名实体识别中的效果。他们发现基于词典的分词信息和分区信息对识别任务有益，并且 SSVM 在其中达到了最好的识别效果。针对社交媒体中的医疗命名实体识别，一些学者也做了相应的研究。为了从社交媒体中提取药物不良反应（adverse drug reactions，ADRs），Nikfarjam 等[5]提出了基于 CRF 的概念提取系统。该模型不仅考虑了上下文信息、ADR 词典、词性、否定等特征，还引入了词向量特征来对词的相似性进行建模。

2 研究方法

与一般文本或电子健康记录不同，网络环境下的医疗命名实体识别存在更多的挑战，如不规范术语、新词语的不断出现等问题。为了提高在线医疗文本的医疗命名实体识别效果，我们提出基于规则和机器学习的综合方法。针对中文在线社区的特点，我们使用中文一体化医学语言系统，并结合网络拓展词典构建了医学领域词典。运用 LDA 模型和语义规则，代替人工标注的过程，最终使用 CRF 模型识别医疗命名实体。实体识别的过程如图 1 所示。它主要包括词典的构建、文本的预处理、分词与词性标注、基于 LDA 和基于规则的命名实体标注、特征标记和实体识别。

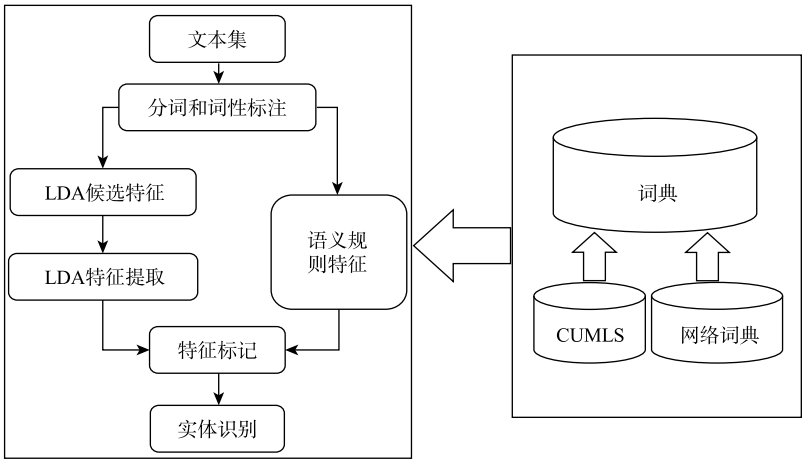

图 1 实体识别过程

2.1 词典构建

对于基于词典的命名实体识别方法，词典的构建至关重要[8]。现有的医疗术语词典多是针对英文环境，适用于中文环境的医疗词典并不多见。本文在医疗实体的识别中采用中文一体化医学语言系统。

它是中国医学科学院医学信息研究所建立的中文医学语料库。CUMLS 由医学词表、语义网、构建工具和平台四个部分构成，整合了十余个生物医学领域内的主题词表、分类表、术语表及医学语料库。医学词表共收录医学主题词 3 万余条、入口词 3 万余条、医学术语 10 万余条、医学词汇素材 30 万余条，医学分类号近万条，融合了自然语言、主题语言和分类语言等情报语言[13]。

尽管 CUMLS 提供了相对全面的词典，但是在线健康社区中的文本内容比较杂乱，表达随意性较强。考虑到 CUMLS 的局限性，我们基于中国两家较为流行的在线健康社区，即寻医问药网（http：//www.xywy.com）和 39 健康网（http：//jbk.39.net/）构建了一个基于网络的健康术语词典。这两个在线健康社区都提供网络上经常使用的医疗概念的解释，我们下载了两个网站上的相关术语构建网络拓展词典。下载的内容分为医疗问题、检查和治疗三个部分。最终，将下载的网络词典和 CUMLS 综合起来构成了本文采用的医疗领域词典，用于接下来命名实体识别的过程。

2.2 分词与词性标注

先对帖子中的文本进行预处理。对无用字符进行过滤，如标点、停用词和特殊字符等。由于汉语没有英文词之间的空格，需要分词处理。鉴于命名实体一般为名词或者名词短语，所以合理的分词和词性标注至关重要。基于构建的医疗领域词典，本文采用 NLPIR 汉语分词系统作为分词和词性标注的工具。

2.3 特征提取

网络上的一些无用信息可能会影响命名实体识别的结果，我们需要提取出那些能代表网络文本主要信息的特征。本文采用 LDA 的主题模型方法提取潜在特征，并采用最短编辑距离[8]的方法决定特征的类型。此外，本文还采用基于语义规则的方法提取和标注某些特征。

2.3.1 基于 LDA 的特征提取

LDA 由 Blei 等[14]在 2003 年提出。它是一种无监督机器学习技术，可以用来识别大规模文档集或语料库中潜在隐藏的主题信息。它的参数不会随着文本集的增加而增加，有很好的泛化能力，是目前机器学习和信息检索领域较为流行的模型。它采用了词袋的方法，将每一篇文档看作一个词频向量，从而将文本信息转化为易于建模的数字信息。

LDA 是一个三层的变参数层次贝叶斯模型。该方法假设一篇文档由一些潜在的主题的多项式分布构成，而每个主题是由不同词的多项式分布组成。LDA 模型描述了文档的生成过程，步骤如下。

（1）对于每个文档 $i \in D$，根据 $\theta_i \sim D_{ir}(\alpha)$，得到多项式分布参数 θ_i。

（2）对于每个主题 $k \in K$，根据 $\psi_k \sim D_{ir}(\beta)$，得到多项式分布参数 ψ_k。

（3）对于文档 i 中的第 j 个词 $W_{i,j}$：①根据多项式分布 $Z_{i,j}$ ~Mult（θ_i），得到主题 $Z_{i,j}$；②根据多项式分布 $W_{i,j}$ ~Mult（ψ_k），得到词 $W_{i,j}$。

在 LDA 模型中，通过迭代推理获得每个文档的主题分布以及相关的主题词的概率分布，使用吉布斯采样对参数（θ，ψ）进行后验推断。如果设置主题数为 K，那么第 i 个词或者短语的概率 w_i 为

$$P(w_i) = \sum_{j=1}^{K} P(w_i | z = j) P(z = j) \tag{1}$$

其中，z 表示一个潜在的主题；$P(z = j)$ 表示 $z = j$ 的概率。

针对某特定主题下的词或者短语的概率分布为

$$P(w_i | z = j) = \{p_{wj1}, p_{wj2}, p_{wj3}, \cdots, p_{wjv}\} \tag{2}$$

其中，$p_{wji}\ (i=1,2,3,\cdots,v)$ 表示在主题 j 下，词或者短语 w_i 的概率。最终得到初始特征为

$$I_1=\left\{w_{ji}\left|p_{wji}\geqslant\varepsilon\right.\right\},j=1,2,3,\cdots,k;\ i=1,2,3,\cdots,v \quad (3)$$

其中，ε 是根据实验决定的参数阈值。LDA 提取主题词的过程中，涉及不同的参数，其中，文档−主题分布 θ_i 和主题−词语分布 ψ_k 在本节中通过吉布斯采样进行估计，而参数 β 采用经验值 0.1，文档的主题数以及每个文档对应的主题词数通过交叉检验，对不同的参数进行取值并取得最优值。最优的主题数 K 由实验的数据得到。在研究中，我们对 β 采用经验值 0.1，α 的取值为 50/K[15, 16]。经过 LDA 主题概率提取之后，得到一系列的主题词，将其作为命名实体识别的候选特征词。

2.3.2 基于语义规则的特征提取

在中文表达中，某些健康命名实体通常伴随着几个固定的字或者词。我们利用这一特点，提出了两种规则来提取在字典中可能不存在的词，作为识别出的特征。首先，通过结合特定医学术语的后缀来识别特征。在中文表达中，人们习惯使用复合词来表示某类疾病或药物。这些复合词通常包含一些常见的后缀，如“病”和“炎”是很多常见疾病的后缀，“片”、“胶囊”和“丸”等是很多常见药物的后缀。因此，我们考虑利用这些独特的后缀来识别特征。如果识别出后缀，则将后缀字和前面的字集组合成为候选特征。其次，我们通过检测一些特定的词前缀来提取特征。中文文本中的一些动词，如“患有”、“诊断为”和“得了”，后面的术语通常都是疾病，表明对自身状况的诊断，即通常表示疾病实体的开始。同时，另外一些动词，如“服用”和“注射”等，后面的词语大都表示药物，即定义药物实体的左边界。这些动词有助于提取特征和识别特征的类型。此外，中文中很多疾病和症状的名称通常与人体部分相关，因此使用从 CUMLS 和网站收集的 329 个身体部位词作为基础，匹配文本中包含身体部位的词以及词组。

2.4 特征标记

特征标记是确定实体边界和类型的过程。本文采用 BIEO（begin-in-end-out）的方式来标记特征，其中，B 类别表示实体的开始；I 类别表示实体的连续，即中间部分；E 类别表示实体的结束；O 类别表示所有非医学实体的其他词或者符号等。本文需要识别的实体类型共有三种，即医疗问题、检查和治疗，标记时分别采用 D、I、T 作为三种不同实体类型的标记，所以共有 10 种不同类型的标记。对于每个识别的特征，特征的第一个词将被标记为 B，并且最后一个词将被标记为 E。特征的中间部分被标记为 I。文档中的所有其他单词都标记为 O。

2.5 基于 CRF 的命名实体识别

CRF 是一种被广泛应用的机器学习方法，在序列标记方面有着出色的表现。在自然语言处理领域，CRF 方法被应用于分词、词性标注和命名实体识别等多项任务中。许多研究证明了 CRF 在命名实体识别上的出色表现[17, 18]。因此，本文选择 CRF 方法对中文虚拟健康社区的医疗命名实体进行识别。

CRF 是由 Laffert 等在 2001 年，结合最大熵模型和隐马尔可夫模型的特点，提出的一种判别式概率无向学习图模型，以给定的输入节点值作为条件来预测输出节点值的概率，是一种用于标注和切分有序数据的条件概率模型[19]。它没有隐马尔可夫模型那样严格的独立性假设。CRF 与最大熵模型的本质区别在于：最大熵模型在每个状态都存在一个概率模型，在状态转移时需要进行归一化，而 CRF 模型在所有的状态上建立了统一的概率模型，能够充分利用上下文信息为特征，同时还能够添加其他的外部特征，使得其能够获得更完善的信息。

一阶链式 CRF 如图 2 所示。

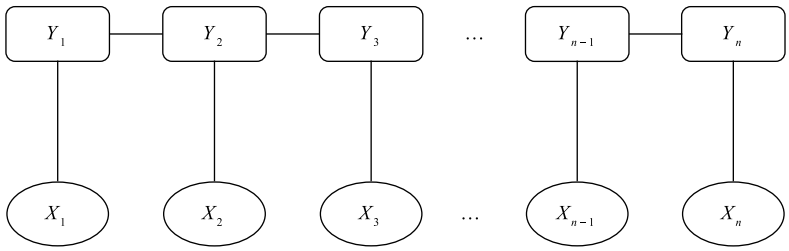

图 2　一阶链式 CRF

令 $X=(x_1,x_2,\cdots,x_n)$ 表示观察序列的集合，$Y=(y_1,y_2,\cdots,y_n)$ 表示状态序列的集合。根据随机场的基本理论，无向图中关于顶点的标记条件概率为

$$P(Y|X,\lambda)=\frac{1}{z(x)}\exp\left(\sum_{i=1}^{n}\sum_{a}\lambda_a f_a(y_{i-1},y_i,x,i)\right) \tag{4}$$

其中，归一化因子

$$z(x)=\sum_{a}\exp\left(\sum_{i=1}^{n}\sum_{a}\lambda_a f_a(y_{i-1},y_i,x,i)\right) \tag{5}$$

以上是状态函数和转移函数的统一表达形式。其中每个 $f_a(y_{i-1},y_i,x,i)$ 是观察序列 x 中位置为 i 和 i−1 的输出节点的特征，λ 是特征函数的权重。分别为训练数据中每一个状态−状态对（y'，y）和状态−观察对（y，x）定义特征，使得建立的 CRF 模型具有类似于隐马尔可夫模型的属性：

$$f_{y',y}(y_u,y_v,x,i)=\begin{cases}1, & \text{如果} y_u=y', y_v=y\\ 0, & \text{其他}\end{cases} \tag{6}$$

3　实验

3.1　数据

我们从家庭医生网站爬取了从2016年2月26日到2016年7月26日的关于糖尿病的咨询问答数据，从中选取有医生回答的帖子作为实验数据，最终获取到共 1 800 条有回答的帖子。将实验数据按照 2∶1 的比例分为训练数据和测试数据，得到 1 200 条帖子作为训练数据，600 条帖子作为测试数据。测试数据由两名研究生作为标注者独立标注出医疗实体及其所属类型，并对标注存在差异的地方进行讨论，最终只保留达成共识的标注。

3.2　指标

对于命名实体识别的结果评估，本文使用准确率（precision，P）、召回率（recall，R）和 F 值（F）作为评估指标。具体而言，准确率是指所有识别出来的命名实体中正确识别的比例；召回率是指所有命名实体中被正确识别出来的比例；F 值是准确率和召回率的调和平均数，可以反映出准确率和召回率的综合表现。P、R 和 F 值的范围为[0，1]，数值越大表示识别效果越好。具体计算方法如下：

$$P=\frac{N_{\text{true}}}{N_{\text{method}}}\times 100\% \tag{7}$$

$$R = \frac{N_{\text{true}}}{N_{\text{total}}} \times 100\% \quad (8)$$

$$F = \frac{2 \times R \times P}{R + P} \times 100\% \quad (9)$$

其中，N_{true} 表示人工标注语料中被命名实体方法正确标注的命名实体数目；N_{method} 表示命名实体识别方法识别出的命名实体数；N_{total} 表示实际正确的命名实体数，即人工标注的正确命名实体数。

3.3 评估

为了评估本文模型的有效性，我们选取了 Xu 等[20]提出的数据驱动方法和 Lei 等[12]使用的基于 CRF 的方法作为比较标准。选择这两种方法的理由有以下几点：第一，这两种方法都是用于中文语料的命名实体识别，可以应用到我们的数据中并进行比较。第二，这两种方法在各自文章语料中都取得了比较好的结果。第三，这两种方法提出的时间较近，而且评价标准相同。第一种方法（Xu 等的方法）使用了混合方法，其中包括构建跨领域核心医疗词典、无监督的迭代算法以及丰富核心词典，应用中文语义规则解决中文书写中的特殊约定。另一种方法（Lei 等的方法）是中文临床文本中命名实体识别的综合研究，通过词典和分词标记，对 CRF 算法以及其他多种命名实体识别方法进行实验分析，其中 CRF 方法有较好的表现。

3.4 结果

首先通过实验确定 LDA 特征提取的最佳参数，最终当主题数量 K 设置为 110 时，使用当前数据集能够达到最大的 F 值。为了避免因测试数据的选择造成的偏差，我们将测试数据随机平均分为 30 组，并将我们的方法和两种基准方法应用在每一个分组测试集中分别计算识别结果。其次使用 T 检验的方法比较本文提出的模型与 Xu 等和 Lei 等的方法，表 1 显示了三种方法在准确率、召回率和 F 值上的均值、标准差。

表 1 实验结果对比

评估指标	P_1（Xu 等）		P_2（Lei 等）		P_3（本文方法）		P_1–P_3（MD）	P_2–P_3（MD）
	Mean	SD	Mean	SD	Mean	SD		
准确率	78.93	5.2	80.54	5.04	79.86	4	−0.93	0.68
召回率	61.97	4.94	60.67	4.4	67.7	3.97	-5.73^{***}	-7.03^{***}
F 值	69.36	4.69	69.14	4.35	73.25	3.69	-3.89^{**}	-4.11^{**}

*表示 $p < 0.1$，**表示 $p < 0.05$，***表示 $p<0.01$

注：Mean 表示平均值；SD 表示标准差；MD 表示平均值差值

如表 1 所示，三种方法的准确率并无明显差异。三种方法都是基于事先构建的词典，所以在识别医疗命名实体的准确性上差别不大。而在召回率和 F 值上，本文提出的方法均优于另外两种基准方法。虽然 Lei 等[12]使用的也是基于 CRF 的方法，但其主要应用于临床文本中的医疗命名实体识别，并没有考虑网络文本的特殊性。基于构建的医疗领域词典及语义规则，本文提出的方法能够成功识别网络环境中部分医疗命名实体的别名和用户产生的新词，从而得到了较高的召回率。这表明本文提出的方法能够识别出更多传统医学词典中不包含的医疗命名实体，更加适用于网络环境下的医疗命名实体识别。

4 结论与讨论

在线健康社区正成为人们交流健康话题越来越重要的场所。随着更多的人在网络中参与健康话题

的讨论，网络健康文本中积累了大量的健康知识，但也造成了人们寻找有价值信息的困难。对于大部分的文本信息处理任务，识别文本中的命名实体往往是第一步。根据识别出来的医疗命名实体，可以进行医疗实体关系的提取，如发现症状与治疗之间、疾病与检查之间和疾病与疾病之间的关系，以及发现药物的不良反应等。这些医疗实体关系的提取可以帮助医生和患者更加科学地进行医疗决策。作为医疗文本挖掘和知识发现的基础，医疗命名实体识别的效果对医疗领域的信息提取任务至关重要。

本文在前人研究的基础上，将医疗命名实体识别从传统结构化的电子病历转移到了网络中非结构化的医疗文本。本文结合网络医疗文本的特点，提出了综合基于规则和机器学习模型的网络医疗命名实体识别方法。该方法基于中文一体化医学语言系统和构建的医疗领域词典，结合 LDA 主题模型和 CRF 模型，识别在线健康社区中的医疗命名实体。实验结果表明，本文提出的方法在召回率和 F 值上均优于两种基准方法。该方法能够广泛应用于医疗领域的数据挖掘与知识发现任务，如医疗实体关系抽取和网络中的药物不良反应的发现等。

本文针对网络环境构建了医疗领域词典，探索了中文在线健康社区中的医疗命名实体识别方法，对文本中的医疗实体按照医疗问题、检查和治疗三部分进行相应的实体识别。本文的创新之处在于以下几点：第一，不同于传统的从电子病历中提取实体，本文试图从在线医疗文本中提取医疗命名实体。第二，考虑到网络医疗文本的特殊性，本文使用了网络拓展词典和基于语义规则的方法，从而提取出传统词典中不包含的医疗命名实体。第三，本文基于 LDA 创建候选特征集，并结合 CRF 方法自动地从文本中提取医疗命名实体，能够适用于大规模的在线医疗文本数据处理任务。

本文尚存在一些不足。首先，本文提出的方法只有在医疗问题和治疗表现上显著优于基准方法，对于检查类命名实体的识别表现并不突出。其次，本文只考虑了以两种近似方法作为对比的基准，未来研究可以考虑以更多样的方法作为比较。最后，本文只考虑了以准确率、召回率和 F 值作为评价标准的情况，并没有考虑算法的计算成本。

参 考 文 献

[1] Kazmer M M，Lustria M L A，Cortese J，et al. Distributed knowledge in an online patient support community：authority and discovery[J]. Journal of the Association for Information Science and Technology，2014，65（7）：1319-1334.

[2] Zhang S，Elhadad N. Unsupervised biomedical named entity recognition：experiments with clinical and biological texts[J]. Journal of Biomedical Informatics，2013，46（6）：1088-1098.

[3] Ning K，Afzal Z，Singh B，et al. Using an ensemble system to improve concept extraction from clinical records[J]. Journal of Biomedical Informatics，2012，45（3）：423-428.

[4] Wu Y，Jiang M，Lei J，et al. Named entity recognition in Chinese clinical text using deep neural network[J]. Studies in Health Technology & Informatics，2015，216：624-628.

[5] Nikfarjam A，Sarker A，O'Connor K，et al. Pharmacovigilance from social media：mining adverse drug reaction mentions using sequence labeling with word embedding cluster features[J]. Journal of the American Medical Informatics Association，2015，22（3）：671-681.

[6] Liu J，Zhao S，Zhang X. An ensemble method for extracting adverse drug events from social media[J]. Artificial Intelligence in Medicine，2016，70：62-76.

[7] Coden A，Savova G，Sominsky I，et al. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model[J]. Journal of Biomedical Informatics，2009，42（5）：937-949.

[8] Song M，Yu H，Han W S. Developing a hybrid dictionary-based bio-entity recognition technique[J]. Bmc Medical Informatics & Decision Making，2015，15（Supp 1）：1-8.

[9] Jiang M，Chen Y，Liu M，et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries[J]. Journal of the American Medical Informatics Association，2011，18（5）：601-606.

[10] Saha S K，Sarkar S，Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition[J]. Journal of Biomedical Informatics，2009，42（5）：905-911.
[11] Sun C，Yi G，Wang X，et al. Rich features based conditional random fields for biological named entities recognition[J]. Computers in Biology & Medicine，2007，37（9）：1327-1333.
[12] Lei J，Tang B，Lu X，et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American Medical Informatics Association，2014，21（5）：808-814.
[13] Danya L，Tiejun H，Junlian L，et al. Construction and application of the Chinese unified medical language system[J]. Journal of Intelligence，2011，30（2）：147-151.
[14] Blei D M，Ng A Y，Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993-1022.
[15] Ma B，Zhang D，Yan Z，et al. An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews[J]. Journal of Electronic Commerce Research，2013，14（4）：304-314.
[16] Griffiths T L，Steyvers M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America，2004，101（Suppl 1）：5228-5235.
[17] Leaman R，Gonzalez G. Banner：an executable survey of advances in biomedical named entity recognition[J]. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing，2008，13：652-663.
[18] Chowdhury F M，Lavelli A. Disease Mention Recognition with Specific Features[C]. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing，2010：83-90.
[19] Lafferty J，McCallum A，Pereira F. Conditional random fields：probabilistic models for segmenting and labeling sequence data[J]. Proceedings of the Eighteenth International Conference on Machine Learning，2001，1：282-289.
[20] Xu D，Zhang M，Zhao T，et al. Data-driven information extraction from Chinese electronic medical records[J]. PLoS One，2014，10（8）：e0136270.

Health-related Named Entity Recognition from Chinese Online Health Communities

YANG Hangzhou[1]，LIU Kai[2]，YAN Zhijun[1]，LI Junlian[3]，SUN Haixia[3]
（1. School of Management and Economics，Beijing Institute of Technology，Beijing 100081，China）
（2. Informatization Office，Beijing Institute of Technology，Beijing 100081，China）
（3. Institute of Medical Information，Chinese Academy of Medical Sciences，Beijing 100020，China）

Abstract Health-related named entity recognition is an essential and fundamental step for knowledge discovering from online unstructured text. Existing approaches have several limits including hard to extract user-created terms and demanding manual annotation. In this work，based on a constructed health domain dictionary，we propose a hybrid model that utilizing Latent Dirichlet Allocation and Conditional Random Field to extract health-related named entities from online health communities. Experimental results indicate that the proposed model outperforms two existing baseline methods.

Key words health-related named entity recognition，online health communities，latent dirichlet allocation，conditional random field

作者简介

杨杭州（1990—），男，北京理工大学管理与经济学院博士研究生，研究方向为管理信息系统、医疗数据挖掘、自然语言处理。E-mail：hangzhou@bit.edu.cn。

刘凯（1992—），男，北京理工大学信息化办公室，硕士毕业于北京理工大学管理与经济学院，研究方向为医疗文本挖掘。E-mail：lkai0101@163.com。

颜志军（1974—），男，北京理工大学管理与经济学院，教授，博士生导师，研究方向为管理信

息系统、医疗信息化、健康管理、电子商务。E-mail：yanzhijun@bit.edu.cn。

李军莲，（1972—），女，中国医学科学院医学信息研究所，研究方向为医学信息学、数字图书馆。E-mail：li.junlian@imicams.ac.cn。

孙海霞，（1984—），女，中国医学科学院医学信息研究所，研究方向为数字图书馆、医学情报学。E-mail：sun.haixia@imicams.ac.cn。