北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2016.020

在线医疗文本中的实体识别研究

苏娅　刘杰†　黄亚楼

南开大学计算机与控制工程学院(软件学院)，天津　300071；† 通信作者，E-mail: nkjieliu@gmail.com

摘要　针对在线医疗文本，设计考虑医疗领域特性的识别特征，并在自建数据集上进行实体识别实验。针对较常见的 5 类疾病：胃炎、肺癌、哮喘、高血压和糖尿病，采用近年来较先进的机器学习模型条件随机场,b进行训练和测试，抽取目标实体包括疾病、症状、药品、治疗方法和检查 5 类。通过采用逐一添加特征的实验方式，验证所提特征的有效性，取得总体 81.26%的准确率和 60.18%召回率，随后对识别特征给出了进一步分析。

关键词　实体识别；数据挖掘；条件随机场；医疗信息

中图分类号　TP391

Entity Recognition Research in Online Medical Texts

SU Ya, LIU Jie†, Huang Yalou

College of Computer and Control Engineering (Software Institute), Nankai University, Tianjin 300071;
† Corresponding author, E-mail: nkjieliu@gmail.com

Abstract　The authors design recognition features with the consideration of medical field characteristic for the online medical text, and the experiment of the entity recognition is carried out on the self-built data set. Concerned about five common diseases: gastritis, lung cancer, asthma, hypertension and diabetes. In the experiment, an advanced machine learning model Conditional Random Field is used for training and testing. The target entities include five kinds: disease, symptoms, drugs, treatment methods and check. The effectiveness of the proposed features is verified by using the experimental method, and the accuracy of the total 81.26% is obtained and the recall rate is 60.18%. Subsequently, the further analysis is given for the recognition features.

Key words　named entity recognition; data mining; conditional random field; medical information

随着生活水平的提高，人们对于健康问题日益关注。互联网行业的迅猛发展催生了一大批在线医疗社区和医疗信息网站，它们为患者提供了多元化的医疗信息获取渠道[1]。这些网站主要以健康知识、疾病信息、医疗新闻等为主要内容，同时也提供用户在线疾病问答功能。在国内，比较知名的有新浪健康、寻医问药、好大夫在线、39 问医生等等。据笔者调查，单是寻医问药网就包含 2004 年 11 月 24 日至今十余年的疾病问答数据，而每天还会涌现数万条新提问。日积月累，这些疾病问答信息将汇成一股非常可观的大数据。这样的数据有着广泛的参与人群，其中包含大量真实的个人案例，潜藏着丰富的医疗价值，然而，它们在文本中大多处于一种非结构化的状态。为了实现信息的充分利用，抽取和挖掘出其中有用的医疗知识，进行命名实体识别通常是第一步。

目前，在医疗领域，针对电子病历、各种医疗报告、医学文献等的实体识别工作已有不少，但针对医疗问答网站中的疾病问答信息尚未见到相关研究，本文即针对这样的问答信息，首次进行实体识

天津市科技支撑项目(13ZCZDGX01098)、天津市自然科学基金(14JCQNJC00600)和中国民航信息技术科研基地开放课题(CAAC-ITRB-201303)资助

收稿日期：2015–06–06；修回日期：2015–08–16；网络出版时间：2015-09-30 11:27:02

别和挖掘工作。本文抽取的实体类别包括疾病、症状、药品、治疗方法和检查五类。在特征选取方面，除了使用一般的实体识别文本特征(例如符号特征、词性特征、英文数字特征等)，还添加了医疗领域特有的一些特征，包括词的后缀特征、身体部位指示词特征来辅助完成识别和抽取工作，最终在自建数据集上达到 81.26%的准确率和 60.18%的召回率。

1 相关工作

命名实体识别是自然语言处理领域一个重要的研究方向。1995 年举行的第六届消息理解会议 MUC-6[2]正式提出命名实体识别任务，它作为文本挖掘中的第一步，主要任务是识别文本中代表其知识主体的词语。MUC 将命名实体主要定义为两类：专有名词和数量词。在不断的研究中，命名实体的含义和范围也在持续地丰富和扩展。MUC 之后，出现了自动内容抽取会议 ACE [3] (Automatic Content Extraction, ACE)，它由美国国家标准技术研究院(NIST)组织创办，从 1999 年开始至今已经举办多次关于信息内容自动抽取的评测任务，ACE 数据集已经成为测试新的信息抽取算法的公认标准。

在生物医学领域，识别对象主要集中在以下几类：电子医疗记录、医学文献和在线医疗社区。目前比较集中的研究是针对医学文献中的基因、蛋白质、药物名、组织名等进行的生物命名实体识别工作[4]。随着医疗系统的信息化，也出现大量针对电子病历进行的识别工作，目前识别 F 值一般在 0.82 左右[5]。

命名实体的识别方法大致包括 3 种：基于词典的方法、基于启发式规则的方法和基于机器学习的方法。基于词典的方法主要通过字符串匹配实现实体识别，但对词典有很强的依赖性。在国外，英文医疗实体识别日趋成熟，可供参考的资料也都比较详实，最著名的词典包括国际疾病分类 ICD-10[6](International Classification of Diseases-10)、医学一体化语言 UMLS[7](Unified Medical Language System)和医学主题词表 MeSH[8](Medical Subject Headings)。在中文方面，国内研究还较少，可供使用的资源也相对匮乏。在基于启发式规则的方法方面，Kraus 等[9]针对大学医疗系统的临床记录，通过构建正则表达式对其中提及的药品、剂量、服用方法等信息进行识别。目前比较流行的是基于机器学习的方法。

命名实体识别，可以看作是一个分类问题，采用类似支持向量机、贝叶斯模型等分类方法；同时，也可以看作是一个序列标注问题，采用隐马尔可夫、最大熵马尔可夫、条件随机场等机器学习方法[10]。Sondhi 等[11]针对医疗论坛 HealthBoards 上的疾病话题信息，利用 SVM 和 CRF 方法进行浅层的信息抽取。在中文方面，浙江大学的叶枫等[12]自建词典，采用条件随机场对电子病历中的疾病、临床症状、手术操作 3 类比较常见的命名实体进行识别，达到 90%以上的 F 值。王世昆等[13]对明清古医案中的症状和病机进行识别，采用 CRF 和 SVM 分别进行训练和测试，这也是在中文方面的较为大胆的尝试。

2 模型和特征选取

在前面提到的众多方法中，条件随机场是一个比较优秀的识别方法，它不仅去除了 HMM 中的独立性假设，而且通过全局的归一化解决了标记偏置的问题，在命名实体识别、词性标注等问题上都取得不错的效果。如果采用 CRF 建立疾病问答中的实体识别模型，将更易于融合新的特征，使用有重叠性非独立的特征，而且利用其强大的推理能力，训练语料中未出现的情况将有可能被识别出来。因此，本文选择 CRF 模型进行医疗文本中命名实体的识别。

2.1 条件随机场模型

条件随机场(Conditional Random Fields, CRF)是一种无向图模型，1958 年由 Luhn 等[14]提出，它提供了一种概率框架，计算在给定一个观察数据序列 $X = (x_1, x_2, \ldots, x_n)$的条件下，该序列所对应标签序列 $Y = (y_1, y_2, \ldots, y_n)$整体出现的概率[15]，即

$$P(Y|X;\theta) = \frac{1}{Z(X;\theta)} \exp\left\{\sum_k \theta_k \psi_k(Y, X)\right\}, \quad (1)$$

$$Z(X;\theta) = \sum_{Y'} \exp\left\{\sum_k \theta_k \psi_k(Y', X)\right\}, \quad (2)$$

其中 $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ 代表模型参数，$\psi_k(Y, X)$ 是任意定义的以 θ_k 为参数关于观察序列 X 和标签序列 Y 的特征函数，$Z(X; \theta)$是归一化因子。

用 CRF 模型进行命名实体识别就可以被视为一个序列标注问题，要识别的每个句子作为一个观

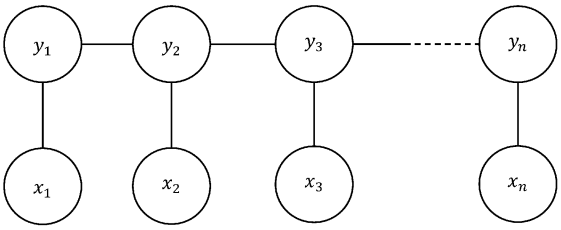

图 1　链式 CRF 结构
Fig. 1　Chain-CRF structure

察序列，句子中的每个词作为一个符号，为每一个符号赋予一个类别标签。CRF 模型最简单的一个结构就是链式结构[16]，如图 1 所示。

进行模型训练时，给定一个训练数据集 $D = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)\}$，其对应经验分布为 $\tilde{p}(X,Y)$，一般可以通过最大化对数似然值，得出模型参数估计:

$$L(\theta) = \sum_{i=1}^{N} \log P(Y_i|X_i;\theta) \propto \sum_{X,Y\tilde{P}} (X,Y) \log P(Y|X;\theta), \quad (3)$$

$$\hat{\theta} = \arg\max_{\theta} L(\theta)。 \quad (4)$$

为了避免过拟合，可以运用一些调整的方法，通常会在参数上加上高斯先验，目标函数 $L(\theta)$就变为

$$L(\theta) = \sum_{i=1}^{N} \log P(Y_i \mid X_i;\theta) - \sum_k \frac{\theta k^2}{2\sigma k^2}, \quad (5)$$

其中σ_k 是高斯先验的方差。得到参数之后，就可以进一步推断给定目标序列 X 最可能的标签序列 $Y*$:

$$Y^* = \arg\max_{Y} P(Y \mid X;\hat{\theta})。 \quad (6)$$

目前已有一些成熟的算法可以用来推断这一值，比如 Viterbi 算法[17]。

2.2　特征选取

对于 CRF 模型来说，特征的选取很关键。通过对疾病问答文本进行分析，本文选择以下特征进行识别。

1) 符号特征。

中文之间没有类似英文空格的天然分隔符，因此在进行实体识别时，需要首先进行分词操作，将分词之后的每一个词语即作为符号特征。为提高分词准确率，引入了自定义词典。通过从多个输入法(包括搜狗、百度、qq)和医疗网站(寻医问药、好大夫在线等)中分别获取，去重合并之后综合成疾病、症状、药物、检查、治疗方法和身体部位词语 6 类辅助词典。

2) 词性特征。

在病人描述中经常会出现“患”、“服用”、“吃”等动词，这些词后会出现疾病名或者药品名，这就为实体的边界的识别提供了线索。在本文中，该特征即为采用 Ansj 分词后的词性。本文采用开源代码库 Github 上的 Ansj[18]系统的分词词性作为这一维特征。

3) 形态特征。

形态特征指当前词的构成情况，包括两个特征:英文字母特征和数字特征。英文字母特征用于标记词当中是否包含有英文字母，因为对于检查来说，经常会出现“ct”、“MRI”之类的英文，而在疾病名、药物等类别中却不常出现。同样，数字特征用于标记该词是否由数字构成。

4) 后缀特征。

在英文命名实体识别中，经常采用词的后缀特征进行识别，而且被证明是有效的。本文研究工作虽然是针对中文开展的命名实体识别，但经观察发现，文本中的各类医疗实体也有一定规律性，比如病名通常以“病”、“症”这类词结尾，而药品则以“颗粒”、“丸”、“剂”等词语结尾，治疗方法则常以“术”结尾。因此，本文也加入后缀特征，即选取词语的最后一个字作为特征。

5) 身体部位指示词特征。

该特征用于标记当前词是否为身体部位相关的词语，因为这样的词语在症状描述中经常出现。

6) 上下文特征。

在词语组成的序列中，上下文之间存在相关性，该特征即为 CRF 模型中的边的特征。当选用不同的窗口长度时，将对各种特征进行组合，形成新的特征。

3　在线医疗文本中的实体识别

针对在线医疗文本信息，我们主要考虑了表 1 中显示的 5 类命名实体。实体识别流程如图 2 所示，主要包括预处理、特征计算、CRF 模型训练和实体识别和识别结果抽取。首先对获取的在线医疗文本进行预处理，包括特殊符号的过滤、人工标注、分词、大小写转化等操作，然后，利用程序从处理好的文本中自动计算并抽取特征，并将所有数据划分为训练集和测试集两部分。将训练集放到模型中进行训练，随后再利用训练得到的参数测试模

表 1　命名实体类别

Table 1　Category of named entity

实体类别	类别定义	标识符号	示例
疾病(Disease)	主要对应于 ICD 中的疾病名、综合征, 概念中可以包括前置的修饰语	D	肺癌、胃炎、糖尿病、咳嗽变异性哮喘
症状(Symptom)	疾病引起的各种不适或异常的表现, 短语中可以包括表示身体部位的词语	S	胸痛、咳嗽、气喘、手脚无力
药品(Medicine)	药物学名、商品名或者通用名称	M	胰岛素、参一胶囊、顺尔宁
治疗方法(Treatment)	与疾病相关的一些手术或疗法	T	姑息术、右肺切除术、四联疗法
医疗检查(Check)	为了确定疾病而采取的检查和测试方法	C	尿常规、血常规、核磁共振、CT 检查

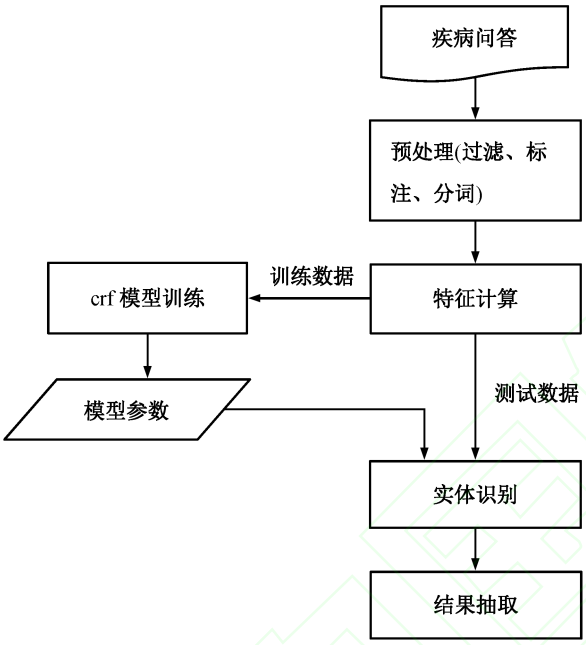

图 2　实体识别流程

Fig. 2　Entity recognition flow

型识别效果。

3.1　数据预处理

为了对在线医疗文本进行实验, 本文采集好大夫在线的五类疾病的全部问答信息, 涉及疾病包括胃炎、肺癌、哮喘、高血压和糖尿病。每一篇文本包含一个提问及相应回答, 其中已经过滤掉没有回答的提问信息。

针对问答文本, 首先进了一些相关的预处理(如对特殊字符、英文大小写、标点符号等的处理)。随后进行了人工数据标注, 采取的标注方式为 BIO 模型(该模型 1995 年由 Ramshaw 等[19]提出), 它可以将分块转化为序列标记确定问题。它的格式为 B-X, I-X 或者 O, 其中 B, I, O 分别代表 Begin, Internal, Other, 即类别的开始、中间或其他, X 代表标注的类别。

识别实验采用了开源工具 CRF++: Yet Another CRF toolkit[20], 其输入有一定的格式要求。标注时首先进行了人工标注, 为力求准确, 对不熟悉的语汇都进行查阅和了解。随后将标注数据转换成所需格式, 如图 3 所示, 前面 6 列分别对应符号特征、词性特征、英文字母特征、数字特征、后缀特征和身体部位指示词特征, 最后一列为标注的实体类别标签。本文采用 Ansj 分词系统, 对于自定义词典中的词都有自定义的类别标签。图 3 中“辅舒酮”为自定义药品词典中的词, 因此词性有别于其他词语。最终对每类疾病分别标注了 200 篇问答信息, 共 1000 篇作为训练和测试数据, 共包含 4812 个不同的实体名词。

crf++采用用户的模板进行特征计算。在选择窗口大小时, 首先在 500 条问答数据上, 采用同样的模板, 设置不同窗口大小进行测试。窗口大小为 3 时的效果优于窗口大小为 1 和 2, 此后再增加窗口大小, 效果提升不大, 因此本文最终将窗口大小设定为 3。针对每一列输入特征 *t*(0 到 5)设置模板,

现在	t	n	n	在	n	O
确诊	v	n	n	诊	n	O
为	p	n	n	为	n	O
哮喘	n	n	n	喘	n	B-D
,	w	n	n	,	n	O
早晚	d	n	n	晚	n	O
吸	v	n	n	吸	n	O
辅舒酮	medicine	n	n	酮	n	B-M
……						

图 3　数据标注示例

Fig. 3　Data annotation sample

包括两类形式:

$$T1 = num: \% \times [index, t], \quad (7)$$

$$T2 = num: \% \times [index, t]/\% \times [index+1, t], \quad (8)$$

其中, num 为模板的编号, index 为窗口大小范围内的索引 (0 到 2), T2 由特征 t 前后位置情况组合而成。

3.2　实验

本文进行两组实验: 第 1 组字典实验验证自定义词典的有效性: 第 2 组为不同特征实验, 通过逐一添加特征的方式, 观察实验效果的变化。实验结果的评测标准由精确度(precision)、召回率(recall)、准确率(Accuracy)和 F1(F1-measure)值构成, 这也是数据挖掘中经常用到的评测指标[21]。对结果的评估采用 conlleval.pl 评测程序[22]。最后针对实验的词性特征和后缀特征进行分析。

3.2.1　字典实验

前面提到, 为提高分词准确率, 自建 6 类医疗词汇词典。然而, 由于是从多个输入法或者医疗网站获取的词汇, 构筑的 6 类词典之间难免会有重叠, 同时其中也充斥着一些不相关的词汇, 由于数量巨大, 如疾病和药品均有上万个词汇, 未能一一过滤, 因此存在噪声。为验证其对实验的影响及添加词典的识别效果, 首先进行字典实验。

实验在一个较小的数据集上开展, 选取 5 类疾病各 100 条问答数据, 采用符号特征和词性特征进行实验。包括 3 组不同的设置, 第 1 组分词时不使用自定义词典, 第 2 组和第 3 组添加同样的自定义词典, 但第 2 组只将自定义词典中的词都标注为同一个词性类别“userDifine”, 而第 3 组根据词语的词典来源标注不同的词性。自定义词典的词数统计信息如表 2 所示。用 B(basic)表示第 1 组, B+D 为第 2 组, D 代表 Dictionary, B+Ds 为第 3 组, Ds 代表多个词典, 不同设置的标注示例如图 4 所示。

实验结果见表3, 可以看出, 添加了自定义词典的识别效果要好于没有添加的情况, 而将词典分为多个不同类别的效果又好于只设定为一个词典的情况。这是因为将词典设置为多个, 相较于设置为一个粒度更细, 因此提供的信息也更为丰富。这组实验也说明虽然词典存在噪声, 但总体来说, 影响不大, 添加多个词典有助于识别效果的提升, 因此, 在下面的实验中, 分词时都将采用多个字典的方式。

表 2　自定义词典情况

Table 2　Condition of user-defined dictionary

类别	总词数
Disease	20562
Medicine	83119
Symptom	466
Treatment	10081
Check	955
Body	270

经常出现	vn	O	经常出现	vn	O	经常出现	vn	O
口腔炎	nhd	B-D	口腔炎	userDefine	B-D	口腔炎	disease	B-D
、	w	O	、	w	O	、	w	O
牙龈炎	n	B-D	牙龈炎	n	B-D	牙龈炎	n	B-D
,	w	O	,	w	O	,	w	O
身体	n	B-S	身体消瘦	userDefine	B-S	身体消瘦	symptom	B-S
消瘦	a	I-S	,	w	O	,	w	O
,	w	O	一直	d	O	一直	d	O
一直	d	O	注射	v	O	注射	v	O
注射	v	O	门冬胰岛素	userDefine	B-M	门冬胰岛素	medicine	B-M
门冬	nw	B-M						
胰岛素	n	I-M						

(a) 不添加词典　(b) 添加词典, 词性均为”userDefine”　(c) 添加词典, 词性为医学专有名词类别

图 4　不同词典设置的标注示例

Fig. 4　Annotation sample of different dictionary settings

表 3　字典实验结果

Table 3　Dictionary experiment results

实体类别	Precision			Recall			F1		
	B	B+D	B+Ds	B	B+D	B+Ds	B	B+D	B+Ds
检查	78.18	71.41	79.51	26.84	26.44	33.42	39.96	38.59	47.06
疾病	66.74	65.43	67.37	18.75	21.42	25.68	29.27	32.27	37.19
药物	77.46	66.79	80.02	50.11	66.29	69.86	60.85	66.54	74.59
症状	70.23	69.10	73.15	28.21	29.58	32.90	40.25	41.42	45.39
治疗方法	72.25	69.02	75.56	23.66	22.33	31.47	35.65	33.74	44.43
总体	74.25	68.48	76.18	30.23	35.14	39.99	42.97	46.44	52.45

3.2.2　不同特征实验

为了验证本文提出的各种特征在问答实体识别中的效果，采用逐一添加特征的方式对 1000 条标注数据进行实验，即每次在符号特征的基础上增加一种特征。首先添加一些常用的实验特征(如词性，英文、数字特征等)，再添加本文提出的后缀和身体部位指示词特征。为了保证实验结果的准确，均采用 5 折交叉验证。图 5 为实验结果总体的变化情况，“word”、“pos”、“al”、“num”、“suffix”、“body”分别代表符号特征、词性特征、英文字母特征、数字特征、后缀特征和身体部位指示词特征。

表 4，5，表 6 和表 7 为实验结果的详细情况。可以看到，随着各类特征的逐一添加，识别精确度略微有下降，主要体现在添加词性特征时，在后面加入后缀和身体部位指示词特征后，精确度又有所回升。总体说来，精确度变化不大。另一方面，实验的召回率在各类实体上都有大幅度的提升，尤其是在药物这一类别最终得到 41.63%的提升，比原来的37.83%的召回率提升了一倍多。F1值在各类实体上也都有了不同程度的提升,总体上,从只用符

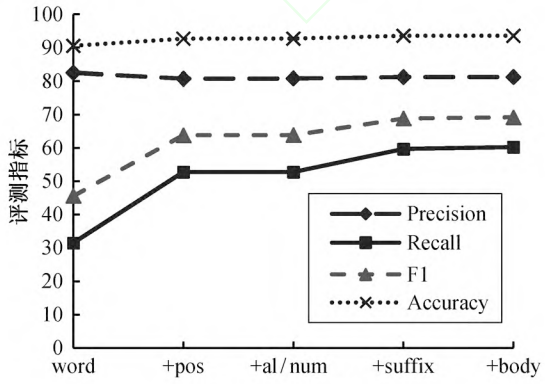

图 5　不同特征实验结果变化情况(总体)

Fig. 5　Experiment result changes with different features (total)

表 4　不同特征实验 Precision

Table 4　Experiment precision result with different features

特征	检查	疾病	药物	症状	治疗方法	总体
word	86.94	79.95	84.72	81.85	74.05	82.48
+pos	84.89	75.73	83.65	78.77	79.49	80.74
+al/num	85.03	75.84	83.56	78.83	79.72	80.80
+suffix	84.00	78.27	85.33	77.76	79.84	81.22
+body	84.08	78.42	85.20	78.04	80.05	81.26

表 5　不同特征实验 Recall

Table 5　Experiment recall result with different features

特征	检查	疾病	药物	症状	治疗方法	总体
word	27.50	27.12	37.83	31.80	26.81	31.43
+pos	45.17	38.88	76.93	46.49	47.64	52.70
+al/num	45.42	38.76	76.79	46.47	47.92	52.71
+suffix	49.61	53.34	79.40	51.30	51.97	59.64
+body	49.93	53.85	79.46	53.00	51.74	60.18

表 6 不同特征实验 F1

Table 6　Experiment F1 result with different features

特征	检查	疾病	药物	症状	治疗方法	总体
word	41.78	40.50	52.30	45.80	39.37	45.52
+pos	58.96	51.38	80.15	58.47	59.58	63.77
+al/num	59.21	51.30	80.03	58.47	59.86	63.80
+suffix	62.38	63.45	82.26	61.82	62.96	68.78
+body	62.65	63.85	82.23	63.13	62.85	69.15

号特征到所有特征都用共提升 23.63%。

实验结果表明，在识别的 5 类实体中，药物的识别效果最好，特别是在召回率和 F1 两个指标上远远超过其他类别的实体。在精确度上，药物最优，其次是检查和治疗方法，最低为疾病和症状，在召回率上正好相反。这可能是因为药名一般比较固定，而且在用户输入信息时格式也比较规整。对于疾病和症状通常有多样化的描述方式，因此识别精度不如其他类别。

表 7　不同特征实验总体 Accuracy

Table 7　Experiment accuracy result with different features

特征	Accuracy
word	90.53
+pos	92.67
+al/num	92.69
+suffix	93.53
+body	93.56

识别结果大致包含以下几种错误类型：1) 识别边界不准确，例如“胸部 ct 检查”只识别出了“ct 检查”，遗漏了相关的指示部位，“中央型肺癌”遗漏了修饰语“中央型”等情况；2) 未识别出较长实体，像“痰中带血丝”、“嗓子老发痒”这样的症状；3) 误分类，例如“腔积液”(疾病)被误分类为药物。导致错误的原因可能与数据集规模有关，下一步可以扩充数据集，丰富特征，寻找真正能抓住其本质的特征进行实验。

3.3　特征分析

对不同特征进行的实验表明词性特征和后缀特征对于识别效果有很大的提升，所以本文进行以下两组分析。

3.3.1　各类实体词性构成模式分析

这里的词性构成综合考虑了当前实体的前一个词的词性、当前词的词性和后一个词的词性，如图 6 所示。

针对被标注为药物的词语“易瑞沙”，分析其词性构成，前一个是动词“服用”，词性为“v”，后一个为标点符号“。”，词性为“w”，当前词词性为药物专有名词“medicine”，因此这个药名的词性构成为“v+medicine+w”。为了对比不同实体类别在词性构成上的情况，我们绘制不同实体的排名前 30 种词性构成模式的频次图，见图 7。可以看出，药物类的曲线非常陡峭，说明药物这类实体的词性的构成在实验文本中是有规律性的，大部分具有固定的模式，因此词性特征才能如此好地提升药物的识别效果。而其他几类，词性也有一些模式，但不如药物明显，因此实验结果也有一定程度的提升。

最近	t	O
开始	v	O
服用	v	O
易瑞沙	medicine	B-M
。	w	O

图 6　词性分析句子示例

Fig. 6　Sentence sample for part of speech analysis

药物类词性构成的前 10 种模式如表 8 所示，可以看出，药物基本上都通过分词被准确标注为药物专有名词，前后出现最多的词性是标点、动词、连词和数词。这也准确反映了文本的潜在结构：用户常将多种药物进行罗列，因此前后出现标点(如顿号、逗号)，出现连词(如“阿法替尼/和/azd9291/效果/怎么样”)；药物名前通常有许多提示性的动词

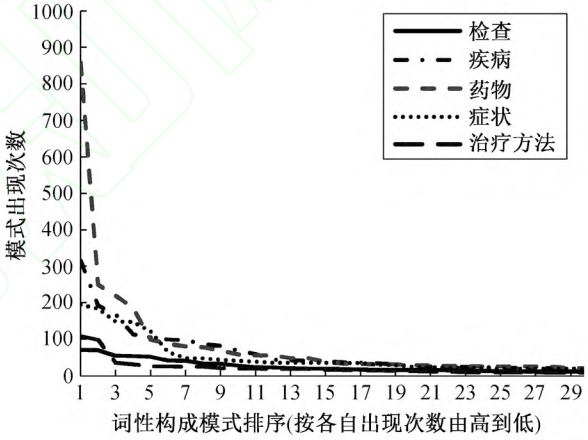

图 7　不同实体词性构成模式频次图

Fig. 7　Different entities pos constitute mode frequency chart

表 8　药物类词性构成前 10 种模式

Table 8　Top 10 patterns of drug class's part of speech

词性模式	频次
w+medicine+w	858
v+medicine+w	249
w+medicine+v	219
w+medicine+m	186
w+n+w	98
v+n+w	88
c+medicine+w	80
v+medicine+m	78
v+n+v	69
v+medicine+v	59

注：w 为标点符号，medicine 为药物专有名词，v 为动词，m 为数词，n 为名词，c 为连词。

(如“服用/口服/使用/注射/吃/开了”); 药物名后, 会紧接着给出服用剂量(如“维生素 c/一天/2/-/3 片”)。

3.3.2 各类实体后缀分析

后缀特征对于疾病类提升较大, 我们统计 1000 条实验数据中疾病名后缀的分布情况, 结果如图 8 所示。在疾病名中,出现最多的前 7 个字分别是炎、病、癌、喘、压、冒、症, 以它们结尾的疾病名共占所有出现的疾病的 64%, 其他 166 个字只占 36%, 说明后缀特征之所以对疾病名称识别有效的原因。

图 9 给出对其他几类实体后缀的分析情况, 可以看到, 不同实体类别的后缀具有不同程度的规律, 因此后缀特征才能有效地提升实验效果。

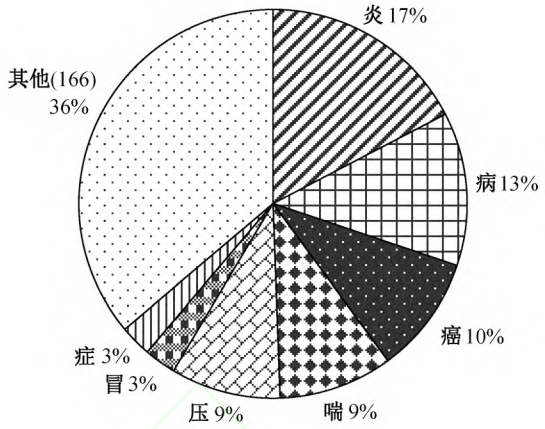

图 8 疾病名后缀分布

Fig. 8 Disease name suffix distribution

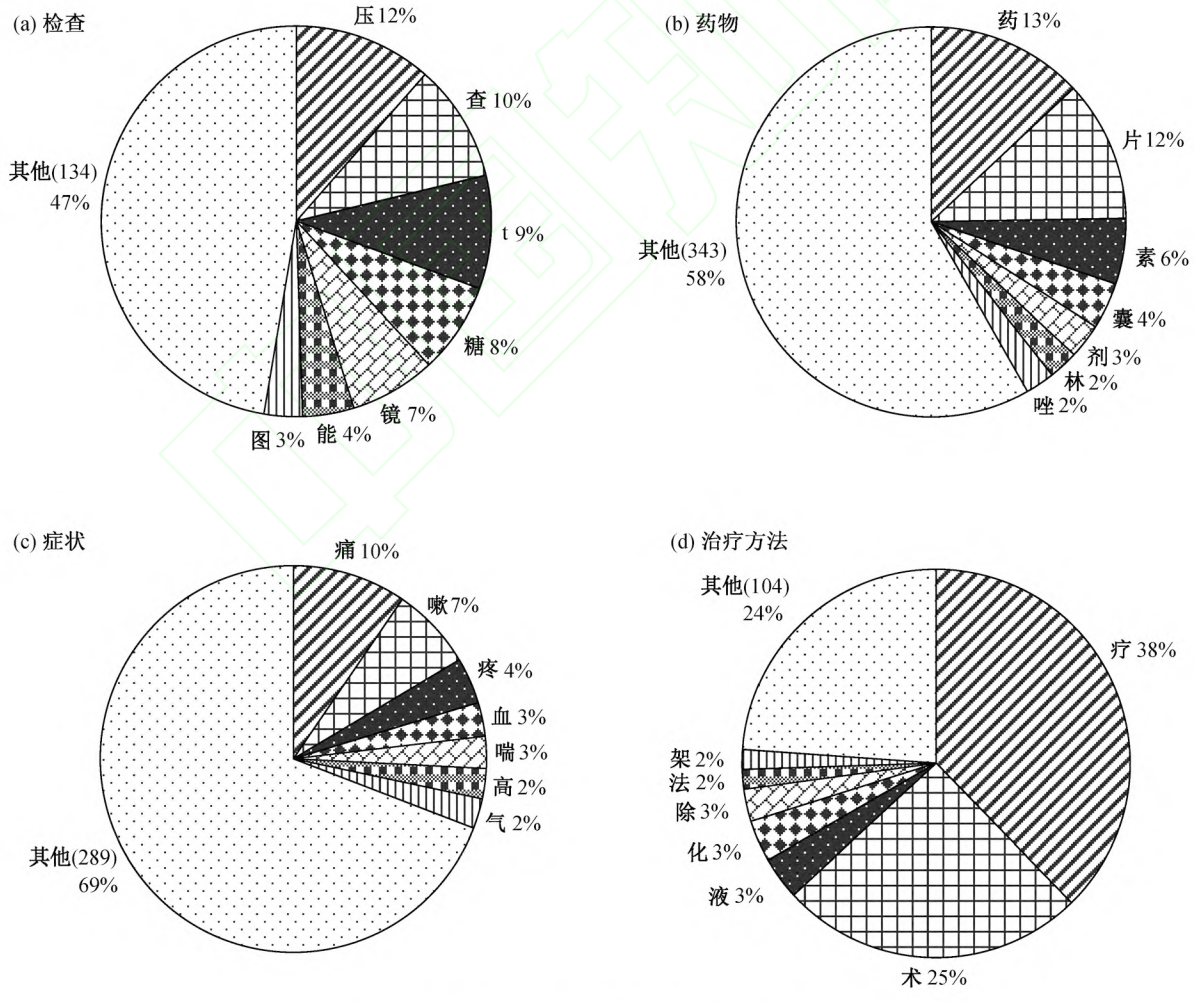

图 9 后缀分布(检查、药物、症状和治疗方法)

Fig. 9 Suffix distribution (check, medicine, symptom and treatment)

4　总结与展望

本文针对在线医疗问答信息，设计了考虑医疗领域特性的识别特征，并采用机器学习模型 CRF 在好大夫问答数据上，针对 5 类医疗实体(疾病、症状、药品、治疗方法和检查)进行了实体识别工作。全部设计特征包括符号特征、词性特征、形态特征、后缀特征、身体部位指示词特征和上下文特征。首先进行了一组字典实验，说明了自定义词典对识别效果的有效提升，然后采用逐一添加特征的方式，观察了实验结果的变化情况。结果表明，随着所提特征的逐一添加，识别精确度有所浮动，而召回率普遍呈现上升趋势，总体的 F1 值也不断上升，当采用所提全部特征时，达到总体 81.26%的精确度和 60.18%的召回率。我们还分析了后缀特征和各类实体的词性构成模式，说明了该特征的有效性。

实验结果表明，本文方法可以有效地识别出问答文本中大部分的医疗实体。但是我们还需继续提高识别准确率，获得更精准的挖掘结果。未来的工作中，我们将进一步丰富实体识别的特征,特别是针对在线的医疗问答文本,进一步区分问与答两种文本的区别和联系，设计相应特征并引入实验；还会考虑前面有否定意义词汇的实体，处理实体嵌套的情况。

参考文献

[1] 黄丹. 网络医疗对医疗服务理念的挑战. 中药研究与信息, 2006, 7(9): 31–32

[2] Grishman R, Sundheim B. Message Understanding Conference-6: a brief history // COLING. Copenhagen, 1996, 96: 466–471

[3] Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ACE) program-tasks, data, and evaluation // LREC. Lisbon, 2004: 837–840

[4] 胡双, 陆涛, 胡建华. 文本挖掘技术在药物研究中的应用. 医学信息学杂志, 2013 (8): 49–53

[5] 杨锦锋, 于秋滨, 关毅, 等. 电子病历命名实体识别和实体关系抽取研究综述. 自动化学报, 2014, 40(8): 1537–1562

[6] DiSantostefano J. International classification of diseases 10th revision (ICD-10). The Journal for Nurse Practitioners, 2009, 5(1): 56–57

[7] Lindberg D A, Humphreys B L, McCray A T. The unified medical language system. Methods of information in Medicine, 1993, 32(4): 281–291

[8] McDonald C J, Overhage J M, Tierney W M, et al. The regenstrief medical record system: a quarter century experience. International journal of medical informatics, 1999, 54(3): 225–253

[9] Kraus S, Blake C, West S L. Information extraction from medical notes // Medinfo 2007. Brisbane, 2007: 1–2

[10] 郑强, 刘齐军, 王正华, 等. 生物医学命名实体识别的研究与进展. 计算机应用研究, 2010, 27(3): 811–816

[11] Sondhi P, Gupta M, Zhai C X, et al. Shallow information extraction from medical forum data // Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics. Beijing, 2010: 1158–1166

[12] 叶枫, 陈莺萤, 周根贵, 等. 电子病历中命名实体的智能识别. 中国生物医学工程学报, 2011, 30(2): 256–262

[13] 王世昆, 李绍滋, 陈彤生. 基于条件随机场的中医命名实体识别. 厦门大学学报: 自然科学版, 2009, 48(3): 359–364

[14] Luhn H P. The automatic creation of literature abstracts. IBM Journal of research and development, 1958, 2(2): 159–165

[15] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data // ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, 2001: 282–289

[16] Sutton C, McCallum A. An introduction to conditional random fields. Machine Learning, 2011, 4(4): 267–373

[17] University of Leeds UK. Hidden Markov Models [EB/OL]. (2010)[2014–11–01]. http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html

[18] 孙建. Ansj seg [EB/OL]. (2012–09–07)[2014–12–01]. https://github.com/NLPchina/ansj seg

[19] Ramshaw L A, Marcus M P. Text chunking using transformation-based learning // Text Speech & Language Technology. Boston, 1995: 82–94

[20] Kudo T. CRF++: Yet another CRF toolkit [EB/OL]. (2005)[2015–03–-01] . http://CRFpp. sourceforge.

net

[21] Powers D M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Mach Learn Technol, 2011, 2(1): 37–63

[22] Tjong Kim Sang E F, Buchholz S. Introduction to the CoNLL-2000 shared task: chunking // Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7. Lisbon, 2000: 127–132