第48卷 第3期 厦门大学学报(自然科学版) Vol. 48 No. 3
2009年5月 Journal of Xiamen University (Natural Science) May 2009

基于条件随机场的中医命名实体识别

王世昆，李绍滋*，陈彤生
(厦门大学信息科学与技术学院，福建 厦门 361005)

摘要：中医医案蕴藏着丰富的知识，如何完成对海量医案的自动标注以便对其进行知识挖掘显得尤为重要.针对明清古医案中症状、病机的自动识别标注问题，采用了基于条件随机场(CRF)的方法，提出数据清洗以及缩减合并词性以减少特征空间规模.最后，通过仿真实验将该方法与最大熵、支持向量机这两种统计方法进行对比.结果表明：该方法在针对明清古医案中症状、病机这类中医命名实体识别具有明显的优势.

关键词：条件随机场；中医命名实体；数据清洗；交叉验证
中图分类号：TP 391.3 **文献标识码**：A **文章编号**：0438-0479(2009)03-0359-06

中医诊断的原理是“司外揣内”，病人表现出来的痛苦不适、神色形态、舌脉变化等，就是认识、揣测内在病理变化的依据.医家们根据这些病理症状通过经验推断病人发病机理，并以此对症下药将整个的诊疗过程以医案的形式记录备份[1].由此我们得出所谓中医医案(medical case)即病案，是指中医治疗疾病时对病人有关的症状、病机、处方、用药等所做的连续记录.所谓的中医命名实体也就是指医案中对病人疾病进行阐述重现的症状、病机、治法、处方、用药等信息实体.这些实体细分下去还有：诊次、病势、转归、预后、医理、医嘱等.

由于医家们在录入医案时主观随意性较大，我们在搜集到的一系列医案语料中经常发现：以上所列中医命名实体的存在是独立而非依赖的，也就是说医案中并未规定某一命名实体出现必定要伴随有另一中医命名实体的出现，既先前阐述的主观随意性.同时我们在整理医案语料的过程中还发现其具有大量不同于其他命名实体如人名、地名甚至生物医学命名实体的特点.由于论文结构上安排的需要，中医命名实体详细特点我们在先引出其他命名实体后再进行阐述.

命名实体识别在新闻领域如：人名、地名、机构名等方面的研究已经获得了很好的效果，F 值评测高达90%以上，已接近人工标注水平.但是目前就我们所掌握的资料还未发现国内有关于中医医案命名实体的相关研究，而对于该类性质的实体研究主要集中在生物医学命名实体之上：如蛋白质、基因、核糖核酸(RNA)、脱氧核糖核酸(DNA)以及细胞的名称等[2]，并且该类命名实体都是针对英文进行.

目前针对生物医学命名实体识别研究准确率和召回率普遍偏低，较新的数据显示，目前识别的准确率大概为70%，召回率为77%左右，F 值约为74%.主要原因有如下几个方面[3]：

1)生物医学命名实体缺乏统一的命名规范，命名主观性强.

2)描述性命名风格导致有的命名实体名称过长.

3)嵌套式结构，即有些生物医学命名实体又包含其他的子实体.

4)连接性实体，及两个实体成并列性构词.

5)随着医学技术的不断进步与发展，新的命名实体不断涌现.

基于以上生物医学命名实体识别的难点，中医命名实体识别不但兼而有之，并且还具有其古汉语特有的问题和难点：医家的常见错别字以及明清古医案中存在着通假字现象和汉语语言独特的歧义词、一词多义、多词一义等比较棘手的文法现象，这使得中医医案的命名实体识别难度加大.

目前，常用的命名实体识别方法有：

1)基于词典的方法——优点是对词典中收录的命名实体具有极高的识别准确率，但由于新命名实体不断出现，并且生物医学实体的命名缺乏统一的规范，所以基于词典的模板匹配对于自由文本效果不佳.

2)基于启发规则的方法——启发规则方法弥补了词典方式不能识别未登录实体的缺陷，使召回率得到明显提高，但带来了准确率降低的缺点，而且人工发现和编写规则比较费时、费力、单调.

3)基于统计的方法——近几年，把基于统计的方

收稿日期：2008-12-10
基金项目：国家自然科学基金(60803078)资助
* 通讯作者：szlig@xmu.edu.cn

法用于命名实体识别渐渐已成为研究的热点. 与基于规则的方法相比,基于统计的方法利用人工标注的语料进行训练,标注语料时不需要广博的语言学知识,并且可以在较短时间内完成,因此这类系统在移植到新的领域时可以不做或少做改动. 此外,基于统计的系统要移植到其他自然语言文本也相对容易一些. 常见的基于统计的命名实体识别方法主要包括隐马尔可夫模型、最大熵模型、支持向量机、决策树以及最新用于NLP 的条件随机场等[4-5].

针对上述命名实体识别方法的优缺点,结合中医医案的特点,本文主要采用统计与规则相结合的方法对中医医案命名实体进行识别研究,在医案语料预处理上采用人工规则进行数据清洗,而在识别方面则采用基于统计的条件随机场进行识别标注,通过仿真实验验证该方法可以获得良好的识别效果.

1 条件随机场(Conditional random fields,CRF)

CRF 是一种无向图模型,可用于最大化条件概率. 常用的特殊图结构是线性链,与一个有限状态机相关,很适合序列标注问题. CRF 可以克服通常的基于有向图的模型标注依赖的问题,且能更好地结合各种信息.

CRF 最早由 Lafferty 等人于 2001 年提出,其思想主要来源于最大熵模型(Max entropy). 我们可以把 CRF 看成是一个无向图模型或马尔可夫随机场,它是一种用来标记和切分序列化数据的统计框架模型. 目前,CRF 在解决英语浅层分析、英文命名实体识别已经取得了良好的效果. McCallum 等人进一步将 CRF 运用到中文分词与新词识别任务中①,其研究成果表明,它能够适用于中文命名实体识别的研究任务.

CRF 是一种无向图模型,假设 X,Y 分别表示需要标记的观察序列和相对应的标记序列的联合分布随机变量,那么 CRF(X,Y) 就是一个以观察序列 X 为条件的无向图模型.

定义 $G=(V,E)$ 为一个无向图,$Y=\{Yv \mid v \in V\}$,即 V 中的每个节点对应于一个随机变量所表示的标记序列的元素 Yv. 如果每个随机变量 Yv 对于 G 遵守马尔可夫属性,即前面所提到的条件独立性,那么 (X,Y) 就构成一个 CRF,而且在给定 X 和所有其他随机变量 $Y_{\{u \mid u \neq v,\{u,v\} \in V\}}$ 的条件下,随机变量 Yv 的概率 $P(Yv \mid X,Yu,u \neq v,\{u,v\} \in V)=P(Yv \mid X,Yu,(u,v) \in E)$.

理论上,图 G 的结构可以是任意的,它描述标记序列中的条件独立性. 但建立模型时,最简单和最普遍的无向图结构是线性链的结构(图 1).

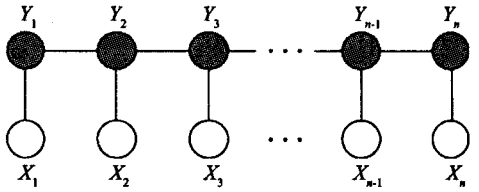

图 1 链状 CRF 的无向图结构

Fig. 1 Undirected graph of chain-like CRF

图中非阴影节点表示的观察值序列并不是由模型产生的. 链状 CRF 假设在各个输出节点之间存在一阶马尔可夫独立性. 需要说明的是,X 的元素间并不存在图的结构,因为我们只是将观察序列作为条件,而并不对 X 做任何的独立假设. 因此链状 CRF 也可以用图 2 表示.

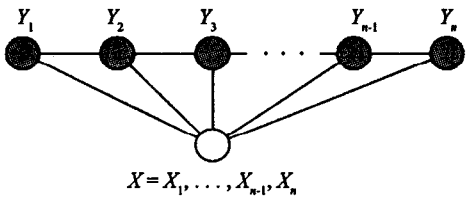

图 2 链状 CRF 的另一种无向图表示

Fig. 2 Another express of undirected graph of chain-like CRF

CRF 中每个状态转移都对应一个非归一化的权值,这意味着在 CRF 模型中的转移是区别对待的. 因此,对任何给定的状态都可能会放大或缩小其传递到后继状态的概率分配,而任意状态序列的权值则由全局归一化因子给出,从而 CRF 也就避免了标记偏置问题的发生①.

假设 O 是一个值可以被观察的输入随机变量集合,S 是一个值能够被模型预测的输出随机变量的集合,且这些输出随机变量之间通过指示依赖关系的无向边所连接. 让 $C(S,O)$ 表示这个图中的团的集合,CRF 将输出随机变量值的条件概率定义为与无向图中各个团的势函数的乘积成正比,即

$$P_A(S \mid O)=\frac{1}{Z_O}\prod_{c \in C(S,O)}\Phi_c(S_c,O_c),$$

其中 $\Phi_c(S_c,O_c)$ 表示团 c 的势函数.

当图形模型中的各输出结点被连接成一条线性链

① Andrew McCallum, Feng Fangfang. Chinese word segmentation with conditional random fields and integrated domain knowledge. Unpublished Manuscript, 2003: 24 – 26.

的特殊情形下,CRF 假设在各个输出结点之间存在一阶马尔可夫独立性,二阶或更高阶的模型可类似扩展. 若让 $O=(O_1,O_2,\cdots,O_T)$ 表示被观察的输入数据序列,让 $S=(S_1,S_2,\cdots,S_T)$ 表示一个状态序列. 在给定一个输入序列的情况下,线性链的 CRF 定义状态序列的条件概率为

$$P_A(S\mid O)=\frac{1}{Z_O}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(S_{t-1},S_t,O,t)\right),$$

其中 $f_k(S_{t-1},S_t,O,t)$ 是一个任意的特征函数,λ_k 是每个特征函数的权值. 归一化因子 Z_O 为

$$Z_O=\sum_{S}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(S_{t-1},S_t,O,t)\right).$$

2　实验验证与分析

本文研究的中医医案原始语料采自明清时期的半文言古医案,实验采用的是 Hieu Xuan Phan 与 Minh Le Nguyen 开发的 FlexCRFs 工具包. 该工具包可在 http://www.jaist.ac.jp/~hieuxuan/flexcrfs/flexcrfs.html 下载[6]. 实验中采用的分词系统是中科院的 ICTCLAS[7],实验的最终目的是对中医医案中的症状和病机这两类命名实体进行良好的识别与标注. 同时本文采用东北大学张乐博士的 Max entropy 工具包[8] 以及 Chang Chih-Chung 等开发的 libsvm-2.84[9] 与 CRF 进行对比性实验验证.

2.1　数据清洗

在使用 ICTCLAS 得到一个粗分词结果后,针对引言中描述的识别需要解决的几个难点问题,在采用 CRF 工具包进行特征训练与测试之前,必须要先对该语料进行相关的数据清洗工作,详细清洗流程如下:

1) 过滤无用词性

针对当时的语言环境,明清医家在记录医案时,常会使用诸如“之”、“乎”、“者”、“也”这类的叹词、助词. 我们认为这些词性对于识别症状、病机不但是无用的,反而会给 CRF 的特征空间带来冗余,因此有必要在数据清洗时将这些词性删除.

2) 校正错别字与通假字

同现代文一样,明清医案也时常会出现一些错别字. 这里提到的错别字并非笔误而是由于语言文化习惯而出现的普遍性错误,如现代文中时常将“走投无路”写成“走头无路”,古文中也常将“神不守舍”写成“神不守色”. 我们所作的清洗工作便是将这类惯常的错别字进行纠正.

通假字也是古文的一大特点,参阅《中医药通假字字典》,常见的通假字诸如:“目”通“木”,“麻目”即“麻木”,“跗”通“浮”,“跗肿”即“浮肿”. 对这些通假字进行校正有助于得出更好的语料模型.

3) 歧义词的切分修正

汉语不同于英语的一个最大特点就是汉语的文字之间没有空格以示区分,而英语中每个单词都是由空格分隔而成的. 这样便造成了汉语中特有的歧义词现象,例如:ICTCLAS 切分“胸腹胀满”的结果为[胸/ng 腹胀/v 满/a],而正确的切分应为[胸腹/ng 胀满/v],再如 ICTCLAS 切分“纳食欲吐”的结果为[纳/v 食欲/n 吐/v],而正确的切分应为[纳/v 食/n 欲/v 吐/v],诸如此类的歧义词还有很多,因此如何对歧义词进行人工干预是一项很重要的前期准备工作.

4) 连接性命名实体的拆分

在引言中曾提到连接性实体是生物医药命名实体识别的一大难点,主要原因便是它使得命名实体过长,识别的最终结果通常忽略了连接词前边的部分. 因此将连接性命名实体进行拆分有助于缩短命名实体长度,提高识别准确率. 如:“膀胱与大肠阻滞”通过初次切分得到[膀胱/n 与/c 大肠/n 阻滞/vn],因此我们只要将连词 c 前后的名词实体膀胱和大肠拆分即可,最终结果为:膀胱阻滞、大肠阻滞.

2.2　特征空间降维

语料清洗后,我们用 CRF 工具包进行训练和测试,发现最终得到的症状识别结果为 77.60%,病机识别结果为 76.33%,效果不甚理想. 经过分析后发现以上的语料模型存在一个比较难以克服的问题,即我们使用的古医案切分工具 ICTCLAS 并未针对古文特征进行过机器学习,因此在对古医案进行切分时造成“单字分割”的情况比较严重. 如:[温 ag/ 补 v/ 不 d/ 效 ng/,,/ 痛 a/ 势 ng/ 日夜 d/ 不 d/ 息 vg/,,/ 饮食 n/ 艰 ag/ 运 v/,,/ 六 m/ 脉 q/ 软弱 a/ 无 v/ 神 n/,,/ 无疑 d/ 虚 d/ 候 v/,,/ 惟 d/ 是 v/ 大 a/ 便 d/ 不 d/ 畅 ag/,,/ 恐 d/ 有 v/ 蓄 v/ 血 n/,,/ 此 r/ 方 d/ 暂 d/ 服 v/..].

以上这例医案除“日夜”、“饮食”、“软弱”、“无疑”这些古现代同义的词语完整切分,其余均零散的切为单字. 考虑到对词性切分的研究不属于本文的重点,我们在分析 ICTCLAS 切分词性标记集时发现,ICTCLAS 词性切分的过于细致,比如:名词方面分为 ng 名语素/n 名词/nr 人名,在形容词方面分为 ag 形语素/a 形容词/ad 副形词/an 名形词,副词方面分为 d 副词/dg 副语素,动词方面分为 v 动词/vg 动语素[10]. 考虑到古文组词造句结构干练,不像现代文结构那么复杂,如果使用的词性过于繁琐,容易使 CRF 特征空间变得冗余,将一些噪音的特征成分加到模型中,令原本就过于零散的分词特征变得更加不具有代表性.

因此基于以上想法,我们将其中的分支词性各选用一种作为代表,即名词 n/ 形容词 a/ 副词 d/ 动词 v,删除其余对古文不太适合的词性. 替换之后原语料变为:[温 a/ 补 v/ 不 d/ 效 n/,,/ 痛 a/ 势 n/ 日夜 d/ 不 d/ 息 v/,,/ 饮食 n/ 艰 a/ 运 v/,,/ 六 m/ 脉 q/ 软弱 a/ 无 v/ 神 n/,,/ 无疑 d/ 虚 d/ 候 v/,,/ 惟 d/ 是 v/ 大 a/ 便 d/ 不 d/ 畅 a/,,/ 恐 d/ 有 v/ 蓄 v/ 血 n/,,/ 此 r/ 方 d/ 暂 d/ 服 v/..].

2.3 模型训练与测试

为了对该模型的性能进行准确评估,防止出现语料库规模过小使得模型学习力度不够,同时也为了防止语料库规模过大而出现训练过拟合的极端情况,采用逐渐递增语料规模的方式找到一个最佳的语料规模. 我们从 500 例医案的语料规模开始,每次递增 200 例医案,其中训练集和测试集所占比例分别为 80% 和 20%. 同时为了验证我们提出的对中医医案进行数据清洗和特征空间降维是有效可行的方案,又设置了一组经过预处理的语料与未经预处理的语料的对比性实验.

每次实验选定语料规模后,为避免选取的语料出现最佳与最差的偶然性事件发生,采用从语料库中随机选取 6 次语料进行交叉验证取平均值的方式. 同时本文采用判别命名实体识别率的一个通用公式

$$F=\frac{(\beta+1)P\times R}{\beta\times P+R} \tag{1}$$

其中 R 为召回率,P 为准确率,β 为召回率和准确率之间的相对权重,通常取 1,故 F 值也称做 F_1 值. 我们通过实验求出的 F_1 值对识别效果进行分析验证. 受篇幅限制,对于不同的语料规模下的模型比较,仅给出 CRF、Max entropy 与 SVM 这 3 种统计学方法其 F_1 值的相应平均结果(见图 3,4),同时在下边的数据分析中考虑到症状和病机的识别在实验中表现出的相似较大,我们仅对症状的实验数据给出分析,最后为了综合比较这 3 种统计学方法的性能,给出了这 3 种方法在各自语料规模下的模型训练时间以供参考(见表 1). 表中 CRF-1、Max entropy-1、SVM-1 是指对未做数据清洗工作和特征空间降维处理的医案语料进行的实验,CRF-2、Max entropy-2、SVM-2 则是指对经过数据清洗和特征空间降维的语料所进行的实验. 实验所用平台为 Windows Vista Ultimate SP1,处理器为 Intel Core2 Duo E4500 2.20 和 2.24 GHz,内存为 2 GB.

通过图 3,4 的症状、病机对比实验图,我们可以明显地看出在经过数据清洗和特征空间降维处理后,这 3 种统计学方法对于中医命名实体的识别率都有了

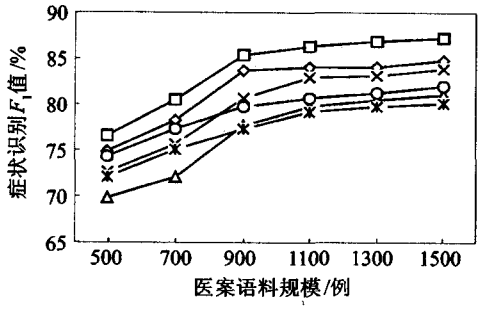

图 3 症状语料规模 F_1 值实验对比图

CRF-1; CRF-2; Max entropy-1; Max entropy-2; SVM-1; SVM-2

Fig. 3 F_1 comparison chart of symptoms corpus

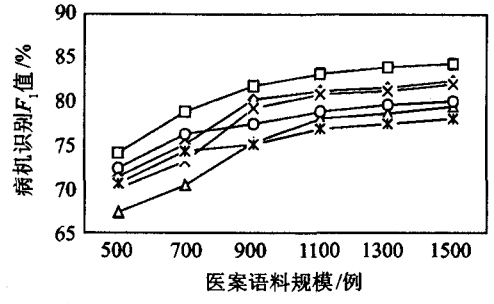

图 4 病机语料规模 F_1 值实验对比图

CRF-1; CRF-2; Max entropy-1; Max entropy-2; SVM-1; SVM-2

Fig. 4 F_1 comparison chart of pathogenesis corpus

不同程度的提高. 由于症状与病机最终的提升特点较为相似,以下仅对症状给予分析.

提升最高比率如下:CRF 在语料规模为 1 300 例时提升了 2.8%,Max entropy 在 1 100 例时提升了 3.15%,SVM 在 900 例时提升了 2.57%;同时也统计了这 3 种方法的平均提升率:CRF 为 2.31%,Max entropy 为 2.85%,SVM 为 1.96%. 通过以上数据对比我们发现 Max entropy 在经过了数据清洗和特征空间降维后提升最为明显,其次为 CRF,最后为 SVM.

同时通过图 3,4 的曲线我们可明显观察到,随着语料规模的不断扩大,无论何种统计模型其识别结果均是稳步上升的. 在最终的症状识别率上,未作处理的语料最好表现依次为:CRF 的 84.75%,Max entropy 的 81.04%,SVM 的 80.12%;而在进行了数据清洗和特征空间降维后依次为:CRF 的 87.16%,Max entropy 的 83.88%,SVM 的 81.92%.

结果表明,无论是否对医案语料进行数据清洗和特征空间降维,CRF 对中医命名实体的识别效果都是最好的,其次为 Max entropy,SVM 则表现的不甚理想. 同时通过图像我们发现在语料规模较小的情况下(实验中语料规模表现为900例以下),SVM的识别效

表 1　症状语料规模训练时间对比表

Tab. 1　Time-consuming comparison table of symptoms corpus

统计方法	训练时间/s					
	1	2	3	4	5	6
CRF-1	478	633	689	777	832	988
CRF-2	292	516	576	694	766	890
Max entropy-1	48	83	123	150	212	326
Max entropy-2	41	71	116	143	205	316
SVM-1	780	2303	3423	4222	5818	7620
SVM-2	590	1312	2716	3433	4988	6027

注：1～6 分别代表语料规模为 500，700，900，1 100，1 300，1 500 例.

果比 Max entropy 要好，但在语料规模较大时 Max entropy 则要强于 SVM.

由于使用这 3 种方法对症状和病机的训练时间相差不大，考虑篇幅限制这里仅给出症状在未经语料处理和经过语料处理两种情况下，这 3 种统计学方法模型训练所需的时间.

再从时间性能上来分析这 3 种统计方法，通过表 1 我们可以明显看出，SVM 的模型训练时间最长，且当语料规模超过 900 例时 SVM 的训练时间均在 1 h 左右，在语料规模最大的 1 500 例时，SVM 训练时间甚至达到了 2 h，远超过 CRF 和 Max entropy. 但是从识别结果来看，高昂的时间代价并未换来好的识别率；Max entropy 模型训练所用的时间最少，通过表 1 可以看出，即使在语料规模最大时，Max entropy 的训练时间都未超过 6 min. 综合图 3，4 的识别率，SVM 除了在语料规模较小时效果好于 Max entropy 外，在语料规模偏大时其识别既费时效果也不尽如人意；CRF 模型训练所需时间适中，基本在 15 min 左右的时间便完成了模型训练，而最终的识别结果也说明了 CRF 在这 3 种统计方法中效果是最佳的.

综上所述，我们认为本文提出的先对医案语料进行数据预清洗，接着再对词性进行合并消解的特征降维，最后采用 CRF 方法进行识别. 这种方法无论从时间复杂度还是最终识别效果来看都是适合中医命名实体识别的.

3　结束语

由上述实验可看出，发现病机的识别效果在准确率、召回率和 F_1 值上都低于症状识别结果. 在查看了具体语料后，我们发现病机由于其抽象性的描述风格（如 [命 v/门 n/火 n/衰 a][温 a/补 v/不 d/效 n][三 m/阴 a/素 d/虚 a]使得其“单字分割”（见 2.2 节）现象比症状的切分要严重，我们认为这些因素客观导致了病机的识别效果要低于症状.

同时，无论是症状还是病机的识别结果都是 Precision＞F_1＞Recall，这说明了我们最终构建的模型在准确率上是令人满意的，也就是说一般由模型标注识别出的命名实体基本上是正确的. 但是在对一些较长的命名实体识别上模型还是有欠缺的地方，在分析了结果后我们发现，对于字数超过 5 个的命名实体，模型一般都无法将其识别，即无法做出正确的召回，这就导致了模型每次在实验时召回率始终低于准确率. 因此，对于如何更好的识别长命名实体是今后工作的一个重点研究方向.

参考文献：

[1] 胡雪琴. 医案语料库的构建及其“内生五邪”病证数据挖掘[D]. 上海：上海中医药大学，2008.

[2] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets[C]//Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. Geneva, Switzerland: COLING, 2004: 104－107.

[3] 胡俊峰，陈蓉，陈源，等. 一种松耦合的生物医学命名实体识别算法[J]. 计算机应用，2007，27(11)：16－19.

[4] 向晓雯. 基于条件随机场的中文命名实体识别[D]. 厦门：厦门大学，2006.

[5] 廖先桃. 中文命名实体识别方法研究[D]. 哈尔滨：哈尔滨工业大学，2006.

[6] Hieuxuan. FlexCRFs: flexible conditional random fields [EB/OL]. http://www.jaist.ac.jp.html.

[7] 中国科学院计算技术研究所. 汉语词法分析工具 ICTCLAS[EB/OL]. http://www.nlp.org.cn/.

[8] Zhang Le. Maximum entropy modeling toolkit for python and C++[EB/OL]. 2007-07. http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.

[9] Chang Chihchung, Lin Chihjen. LIBSVM —— a library for support vector machines[EB/OL]. http://www.csie.

ntu. edu. tw/～cjlin/libsvm.

[10] 俞士汶,朱学锋,段惠明. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报,2002,16(5):49－64.

Recognition of Chinese Medicine Named Entity Based on Condition Random Field

WANG Shi-kun, LI Shao-zi*, CHEN Tong-sheng

(School of Information Science and Technology, Xiamen University, Xiamen 361005, China)

Abstract: Traditional Chinese medicine contains rich knowledge, how to complete the medical case of mass tagging in order to extract their knowledge seems particularly important. This paper uses CRF to mark symptoms and pathogenesis in Medical Records of the Ming and Qing dynasties. In accord with characteristics of chinese we put forward a proposal of data cleansing, and we combine the part of speech in order to reduce the size of feature space. In order to verify the superiority of CRF in Chinese medicine named entity identification, we use maximum entropy and SVM to compare with CRF. The results showed that: our methods proposed for the Ming and Qing dynasties in the case of ancient medical symptoms, pathogenesis such Named Entity Recognition of Chinese medicine has a distinct advantage.

Key words: condition random field (CRF); traditional Chinese medicine named entity; data cleansing; cross validation