电子病历文本症状自动识别方法*

龚 凡 王梦婕 阮 彤 王昊奋 陆 灏

（上海中医药大学附属曙光医院 上海 201203） （华东理工大学 上海 200237） （上海中医院大学附属曙光医院 上海 201203）

〔摘要〕 基于症状体系识别的难点，提出一种创新的基于症状构成模式的非监督学习方法来实现电子病历症状实体的自动抽取，介绍其总体过程并与基于 CRF 序列标注的监督学习方法进行比较，试验证明本文所提出的方法具有良好的识别效果和可扩展性。

〔关键词〕 医疗实体抽取；症状构成模式；结构化电子病历

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.07.002

Automatic Recognition Methods of Symptoms in Texts of Electronic Medical Records *GONG Fan, Shuguang Hospital Affiliated to Shanghai TCM University, Shanghai 201203, China; WANG Meng－jie, RUAN Tong, WANG Hao－fen, East China University of Science and Technology, Shanghai 200237, China; LU Hao, Shuguang Hospital Affiliated to Shanghai TCM University, Shanghai 201203, China*

〔**Abstract**〕 Based on difficulties in the recognition of the symptom system, the paper proposes an innovative unsupervised learning method based on the forming mode of symptoms so as to realize automatic extraction of symptomatic entities of Electronic Medical Records (EMR). It introduces the overall process, and compares this method with the supervised learning method based on CRF sequence labeling. The test proves that the method proposed in this poper has better recognition effect and extendibility.

〔**Keywords**〕 Medical entity extraction; Symptom composition pattern; Structured Electronic Medical Records

1 引言

信息技术的迅猛发展带动了医院的信息化建设，国家政策的支持为电子病历系统等医学信息系统的建立打下了坚实的基础；由此带来了大量的医疗数据，而其中电子病历受到了广泛的关注。电子病历是在医疗活动过程中产生的重要临床信息资源，包含大量与患者健康状况密切相关的医疗知识，从纷繁复杂的电子病历中挖掘出有价值的信息将大大推动医疗事业的发展。电子病历文本结构化是医疗数据挖掘的第 1 步，而其中症状实体识别又是病历结构化的难点，原因有二：一是症状实体的抽取没有标准症状库可以利用，针对疾病实体有 ICD[1] 编码、LOINC 编码[2]、医学主题词表 MeSH[3]、临床医疗术语集 SNOMED－CT[4]、一体化医学语言系统 UMLS[5] 等多种分类体系；二是病

〔修回日期〕 2016－06－02

〔作者简介〕 龚凡，在读博士研究生，发表论文 3 篇；通讯作者：陆灏。

〔基金项目〕 上海市中医药事业发展三年行动计划（项目编号：ZY3－CCCX－2－1003）；国家高技术研究发展计划“心血管疾病与肿瘤疾病中西医临床大数据处理分析与应用研究”（项目编号：2015AA020107）。

历文本在描述患者所患症状时，常用程度、性质甚至否定等修饰词，使得症状的表述形式多种多样。

基于症状实体识别的难点，本文提出基于症状构成模式的方法进行中文电子病历的症状抽取。该方法不同于抽取医疗实体常用的监督学习方法，无须大量的人工标注工作；而且，该方法所用词库还可以根据需要进行扩充，具有良好的可扩展性。本文着重于西医症状实体的抽取，以阐述方法的有效性；对于中医症状实体的抽取，可采用类似的方法。

2 医疗实体抽取的相关工作

2.1 基于语言规则的抽取

许华等[6]基于人工总结规则抽取了药品说明书中的疾病、症状和致病菌 3 类实体，但其并未对抽取症状实体所用的规则进行说明，从给出的抽取致病菌实体的规则示例可知，人工定义的规则过于简单，不适用于表述复杂的症状实体的抽取。国外也有类似基于语言规则的抽取方法，称为 Hearst Pattern[7]，通常与远程监督（Distant Supervision）方法[8]结合进行信息抽取，如基于模式（Pattern）[9]远程标注医疗信息三元组。

2.2 基于监督学习方法的抽取

监督学习方法（Supervised Learning）是命名实体识别中较为常用的方法。对于医疗实体的抽取，通常将其视为序列标注问题，使用条件随机场（Conditional Random Fields，CRF）模型抽取病历文本中的症状、诱因、疾病 3 类实体[10]。同样也是基于 CRF 模型识别电子病历中的疾病、症状、手术操作 3 类实体[11]。Wang[12]等对监督学习方法识别中医症状实体进行了较为系统的研究，其使用 N - gram 特征和位置特征等四维特征，训练并对比了隐马尔科夫模型（Hidden Markov Model，HMM）、CRF 和最大熵隐马尔可夫模型（Maximum Entropy Markov Model，MEMM）3 个经典序列标注模型，试验表明 CRF 模型的结果最优。

2.3 基于增量迭代学习方法的抽取

徐永东等[13]设计了病历信息的五元组表示：<对象修饰，对象，程度，性质，对象描述>，获取种子集到病历文本中学习出实例对应的模式，然后利用学习到的模式继续到病历文本中匹配和抽取病历信息并添加到种子集中，以此增量迭代；但其仅用了少量种子（仅 29 个种子）抽取 < 对象，对象描述 > 二元关系（如 < 皮肤，肿胀 >），不能代表增量迭代学习方法应用在中文医疗信息抽取时的效果。国外已有许多工作将增量迭代方法应用于医疗实体的抽取。如利用现有症状集基于工具 SDMC extractor 从文本语料中学习出模式，然后应用习得的模式从文本中抽取出新的症状实体加入到症状集中，不断增量迭代[14]。Okumura 等[15]使用医疗命名实体识别工具 MetaMap[16]对文本语料标注 UMLS[17]标签，对这些标签进行分类，然后找出与症状相关的分类；同时，还总结了一些症状实体构成的启发式信息用于组合 UMLS 标签以进一步抽取症状实体。此外，Bing 等使用远程监督和增量迭代相结合的方法抽取了医疗数据中的实体[18]和实体关系[19]，首先利用 Freebase[20]中的数据远程标注医疗文本数据，然后利用文本中包含并列关系的句子构建二部图，在二部图上进行标签传播，用于确定包含歧义的实体对应的类型，由此识别出文本中不同类型的医疗实体。

3 基于症状构成模式的电子病历症状识别方法总体流程

3.1 概述

症状实体的抽取不同于疾病、检查等医疗实体的抽取，没有标准的症状库可以利用；同时，症状实体本身具有表述形式多样的特性，如描述“头痛”可以用“头痛剧烈”、“明显头痛”等表达，无法构建一个包含所有表达的完整症状库。但是，仔细观察表现，症状通常可由最基础的症状词到包含否定词、修饰词、部位词、症状词的模式组成，如症状“紧张性头痛”可由模式“修饰词（紧张性）+部位词（头）+症状词（痛）”组成。基于这种启发，本文提出基于症状构成模式的方法自动

识别病历文本中的症状实体。基于症状构成模式的症状实体识别方法总体流程，见图 1。对于输入的病历文本，首先基于西医症状词库对文本进行分词，识别出西医症状实体在文本中出现的位置；然后基于模式对识别结果进行后处理，确定病历文本中西医症状实体的边界；最后输出该病历文本包含的西医症状列表。由图可知，“识别症状实体位置”步骤需要西医症状词库实现基于词典分词的方法，“确定症状实体边界”步骤需要修饰词词典、否定词词典和部位词词典完成后处理。

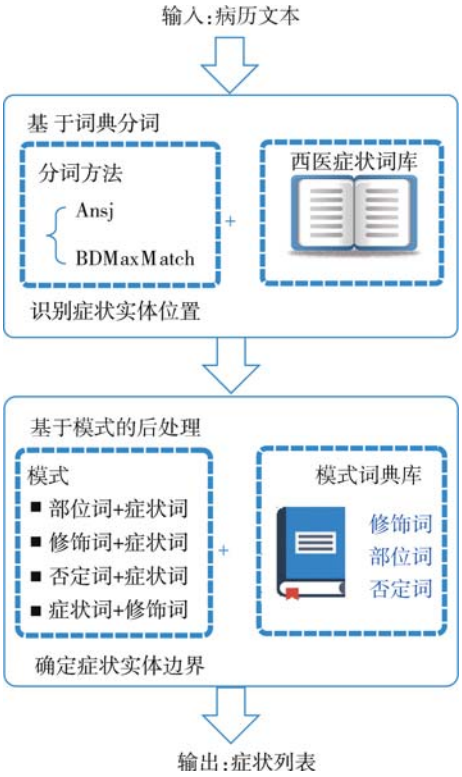

图 1 基于症状构成模式的电子病历症状识别方法总体流程

3.2 识别症状实体位置

3.2.1 *西医症状词库的构建* 为了识别电子病历文本中西医症状实体出现的位置，需构建西医症状词库。词库包含两部分数据：基于医疗领域网站爬取得到的医疗实体子集和基于论文[21]中的 SegPhrase 方法对病历数据进行处理收集得到的西医症状子集，构建过程，见图 2。

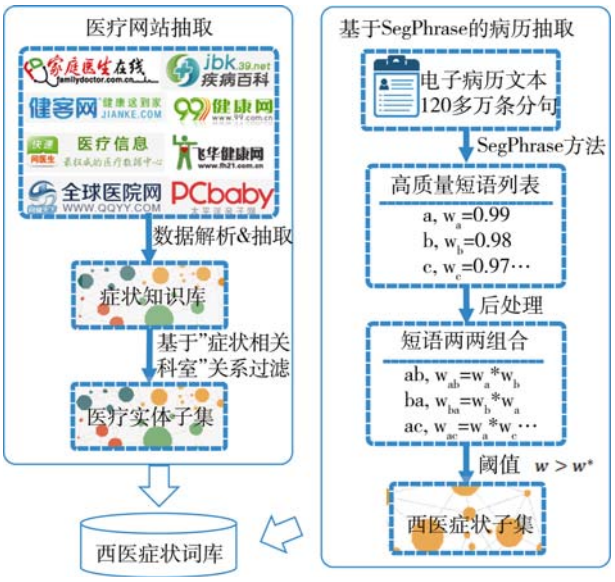

图 2 西医症状词库的构建过程

首先，基于网页的 HTML 标签特性对表 1 中 8 个医疗网站下的症状实体页面进行解析和抽取。以“99 健康网”中症状“关节疼痛”页面为例，见图 3，可以解析的内容包括症状名称、症状部位、相关科室和相关疾病，对应本文定义的类型、症状相关部位、症状相关科室和症状相关疾病 4 种关系。本文对这些页面抽取 < 主语，关系，宾语 > 三元组，如 < 关节疼痛，类型，症状 >，最终构造了包含 16 种关系、38 160 个医疗实体和 367 524 条三元组的症状知识库。鉴于医疗网站包含中医症状和西医症状，使用症状知识库中的“症状相关科室”关系过滤掉中医症状实体，即如果一个症状来自中医相关科室，则认为该症状实体是中医症状。基于此方法收集了 10 646 个西医症状实体；同时，从中随机抽出 300 个西医症状进行人工评测，结果均为西医症状，证明了本文所收集的数据有效可用。

表 1 爬取的医疗网站

名称	网址	包含症状数量(个)
飞华健康网	http://www.fh21.com.cn/	10 180
家庭医生在线	http://www.familydoctor.com.cn/	5 240
全球医院网	http://zz.qqyy.com/	4 260
太平洋亲子网	http://www.pcbaby.com.cn/	1 826
39 健康网	http://www.39.net/	6 675
99 健康网	http://www.99.com.cn/	2 117
快速问医生	http://www.120ask.com/	7 691
健客网	http://www.jianke.com/	8 097

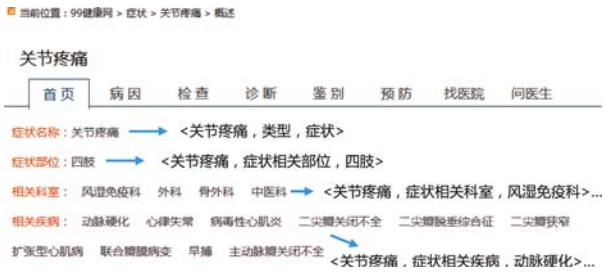

图 3 “99 健康网”的症状实体页面示例

需要说明的是本文将收集的西医症状实体集和症状知识库中其他类型的医疗实体一起作为医疗实体子集添加到西医症状词库中，分配特殊的词性标识类型：症状、疾病、药物、科室、部位和检查，以此减少医疗实体的歧义性。此外，本文对从上海中医药大学附属曙光医院内分泌科得到的 16 438 份电子病历文本进行分句，使用基于 SegPhrase 的方法学习出西医症状实体，分配词性 sympTerm（症状），添加至西医症状词库。SegPhrase 方法主要面向大规模的同领域文本语料总结出代表该领域的高质量短语列表，该方法通过四维特征度量短语的质量：（1）普遍性——高质量短语在语料中应具有高频率。（2）语义一致性——高质量短语的频率应远高于随机组合短语的频率，如 strong tea（浓茶）要比 powerful tea（无意义的组合）的频率高。（3）信息性——高质量短语应和特定主题相关，如“本文”一词出现的频率很高，但并不具有信息性。（4）完整性——一个短语和其子串可能都满足以上 3 个特征，但高质量短语在特定的上下文中应作为一个完整的语义单元，如症状“肠鸣音亢进”是一个具有完整性的高质量短语，而短语“肠鸣音”并不具有完整性。

SegPhrase 方法通过 5 个步骤抽取语料中的高质量短语：（1）统计语料中所有短语的出现频率，将满足频率阈值的短语加入短语候选集，以此满足普遍性特征。（2）人工标注极少量的数据作为训练集，基于语义一致性和信息性相关的特征（如 PMI、IDF）训练得到随机森林（Random Forest）模型；使用该模型评估候选短语的质量，输出 0 ~ 1 之间的权重 w。（3）通过相关文章[14]提出的“短语分词模型”对每个候选短语的出现频率进行修正，称为“修正频率”，以满足完整性特征。（4）基于修正频率得到二维分词特征，加入原特征集中，训练出一个新的短语质量评估模型；重复步骤（2）和（3）。（5）过滤掉具有低修正频率的短语，输出高质量短语列表（按照短语权重排序）。

如图 2 所示，通过 SegPhrase 方法对电子病历文本进行处理，可以得到高质量短语列表。试验发现，SegPhrase 方法得到的短语最大长度为 6，而观察电子病历文本可知，症状实体大多包含部位词和修饰词，短语长度超过 6，这说明 SegPhrase 方法得到的高质量短语列表没有覆盖到这类较长的症状实体。因此，本文对高质量短语列表进一步处理，将其两两组合，形成新的短语。如短语“自觉乏力明显”和“好转”（权重分别为 0.58 和 0.59），组成新的短语“自觉乏力明显好转”，且出现在病历文本中，就将该短语加到高质量短语列表中，其权重是两个短语的权重乘积，即 $w = 0.58 \times 0.59 \approx 0.34$。如果出现多个短语组合结果一致的情况，则取权重最高值作为合并后短语的权重。最后，对后处理的高质量短语划分阈值，添加到症状词库中。

本文标注了 200 个短语作为训练集，在电子病历语料上通过 SegPhrase 方法得到了 13 804 个短语，经过后处理得到 27 142 个短语，最终收集了 400 个西医症状词。通过以上两种方法，完成了西医症状词库的构建。

3.2.2 *基于词典分词* 本文采用两种基于词典的分词方法来识别病历文本中症状实体的位置：双向最大匹配算法（Bi - Direction Maximum Matching，BDMaxMatch）和分词工具 Ansj[22]方法。

BDMaxMatch 算法是一种常见的基于词典的机械分词方法，简单而言，就是其综合了前向最大匹配（Forward Maximum Matching，FMM）算法和后向最大匹配（Reverse Maximum Matching，RMM）算法的结果。FMM 算法的大致流程为：记词典中最长词语长度为 MaxLength，在待分词字符串中，自左向右取长度为 MaxLength 的子字符串，与词典进行匹配。如果词典中包含该子字符串，则将其分出来，向后再取 MaxLength 长度的子字符串与词典进行匹配；否则，将取出的子字符串去掉最后一个字后与词典匹配。以此方式重复，直至将原字符串处理完。RMM 算法则从待分词字符串的最后一个字开始，从后往前取子字符串，其匹配逻辑和 FMM 算法相同。BDMaxMatch 是将待分词字符串分别用 FMM 和 RMM 进行分词，然后计算两种分词结果中

单字的个数，取单字个数较少的结果作为其结果；如果单字个数相同，则取包含短语个数多的结果，否则，取 RMM 的结果作为其结果。

本文在使用 BDMaxMatch 时，所用词典包含西医症状词库和模式词典库，分别用来切分出电子病历中的症状词和症状的修饰词。另一种基于词典分词的方法借助了工具 Ansj 来实现分词。本文使用了 Ansj 的 NLP 分词方式，具有用户自定义词典、新词发现等功能。在借助 Ansj 对病历文本进行分词时，需将西医症状词库加入用户自定义词典。为使 Ansj 在分词时优先考虑西医症状词库中的词，在将西医症状词库加入用户自定义词典时，症状词的权重均设置为最大值。以上两种方法分别对电子病历文本进行分词后，重新遍历病历文本，若当前词的词性为 sympTerm，就将其作为症状实体出现的位置，由此完成了症状实体位置的识别。

3.3 确定症状实体边界

3.3.1 *模式词典库的构建* 症状实体通常有几种固定的构成模式，简单如“腰酸”、“背痛”是由“部位词（腰、背）+症状词（酸、痛）”模式组成，复杂如病历中描述的“浅表淋巴结未及明显肿大”是由“部位词（浅表淋巴结）+否定词（未及）+修饰词（明显）+症状词（肿大）”模式组成。通过观察大量病历文本并和领域专家探讨，本文总结了症状实体的一般构成模式，由此形成基础模式集合，见表 2。

表 2 症状构成的基础模式

序号	基础模式
1	部位词 + 症状词→症状词
2	修饰词 + 症状词→症状词
3	症状词 + 修饰词→症状词
4	否定词 + 症状词→症状词

基于基础模式扩展得到构成复杂的症状实体，如症状“浅表淋巴结未及明显肿大”的扩展过程为：修饰词（明显）+症状词（肿大）→否定词（未及）+症状词（明显肿大）→部位词（浅表淋巴结）+症状词（未及明显肿大）→症状词（浅表淋巴结未及明显肿大）。基础模式涉及的症状词即在“识别症状实体位置”过程中得到的标识症状实体位置的短语。而基础模式所需的部位词、修饰词和否定词则需要分别收集。通过数据观察发现，位于症状词前的修饰词和位于症状词后的修饰词有所区别，因此本文对基础模式“修饰词 + 症状词→症状词”和“症状词 + 修饰词→症状词”涉及的修饰词分别进行收集，这里将它们分别称为前向修饰词和后向修饰词。由此，模式词典库包含 4 类词典：部位词词典、前向修饰词词典、后向修饰词词典和否定词词典。部位词词典来源于构建西医症状词库过程中得到的症状知识库。在解析和抽取医疗网站中的症状页面时，会得到关系为“症状相关部位”的三元组，基于此收集“类型”为“部位”的实体，便构成了部位词词典。

前向修饰词词典和后向修饰词词典的构建方法类似，均是使用分词工具 Ansj 实现的“基于词典分词”方法对电子病历文本语料的所有分句进行分词，找到标识症状实体位置的短语，然后将该短语前后两个词取出来分别加入前向修饰词候选集和后向修饰词候选集。下一步对两个候选集中的词语按照词频进行排序，分别取候选集中的高频词加入修饰词词典（基于对候选集的观察，本文将前向修饰词词典和后向修饰词词典的频率阈值分别设为 90、199）。最后，本文收集了电子病历中常见的否定词（如“无”、“未及”、“没有”）加入否定词词典。经过以上几步，完成了模式词典库的构建。

3.3.2 *基于模式的后处理* 即从病历文本中标识症状实体位置的短语出发，基于确定症状实体边界，由此抽取出症状实体。为了识别症状实体的边界，使用如下识别策略：从症状词（标识了症状实体在病历文本中的位置）出发，向前查找部位词、前向修饰词和否定词，向后查找后向修饰词，将满足条件的短语合并至症状词中，直至遇到不满足条件的短语，停止合并，并输出合并后的症状词，作为识别出的症状实体。这一过程用伪代码描述如下：

Algorithm 1 基于模式的后处理方法

Input：（1）经过基于词典分词方法处理后的文本 phraseList

（2）标识 phraseList 中症状实体的症状词 sympTerm

（3）模式词典库：部位词词典 positionDict、前向修饰词词典 preModifierDict、后向修饰词词典 postModifierDict、否定词词典 negativeDict

Output：symptomList

1：symptom←sympTerm；

2：preposition←sympTerm. position - 1；

```
3：postpostion←sympTerm. position + 1；
4：while（phraseList. get（preposition）in positionDict, preModifierDict or negativeDict）｛
5：//向前匹配：症状词前面的词与部位词、前向修饰词和否定词词典进行比较
6：symptom←phraseList. get（prepotion） + symtom；//满足条件的词与症状词合并
7：preposition - -；
8：｝
9：while（phraseList. get（postposition）in postModifierDict）｛//向后匹配
10：symptom←symtom + phraseList. get（postpotion）；
11：postposition + +；
12：｝
13：symptomList. add（symotom）；
14：return symptomList；
```

完成基于模式的后处理工作，就确定了病历文本中症状实体的边界，由此实现了症状实体的自动识别。

4 试验结果和分析

4.1 语料和症状标注范围

试验的数据源为上海中医药大学附属曙光医院内分泌科的 16 438 份电子病历文本。本文从中抽取出 60 份患者的首次病程记录（包含入院概况、现病史、体格检查、西医检查、入院诊断和治疗方案 6 方面信息），人工标注了文本中出现的症状实体，作为试验语料。其中，40 份作为训练集，剩下 20 份作为测试集。本文仅标注病历文本中的西医症状，标注范围包括：（1）异常症状（异常的体征也包含在内），如症状“自觉乏力明显”、体征“四肢活动不利”。（2）用否定词表示的“正常症状”，如“无腹痛”、“皮肤黏膜无黄染”。这里之所以将“否定词表示的正常症状”纳入标注范围，是从病历文本中出现的症状一般与患者所患疾病相关，“患有某症状”和“不患有某症状”同等重要的角度出发。

4.2 SegPhrase 方法的阈值选取

本文对试验所用数据源进行分句处理，共得到 1 274 571 条分句，然后使用 SegPhrase 方法学习出西医症状子集。如前文所述，SegPhrase 方法需要标注极少量的数据作为训练集。相关文献[21]中说明了标注 200 ~ 300 个数据即可训练出有效的模型，同时需要标注正例和负例。针对本文的抽取目标，从西医症状词库中随机抽出 100 个字符串长度不大于 6 的症状词作为训练集的正例，然后从 SegPhrase 在第 1 步生成的短语候选集中随机取 100 个短语进行标注，其中将完整的西医症状词作为正例，其他词为负例，由此生成 SegPhrase 方法所需训练集。将病历文本分句和训练集作为输入，使用 3.2.1 节所述方法对语料基于 SegPhrase 方法学习出高质量短语列表，进行后处理，最终得到 27 142 个高质量短语。本文标注出高质量短语列表中出现西医症状实体，进一步计算出不同权重区间内西医症状实体所占比例，以确定阈值。结果见图 4，横轴为权重区间（差值取 0.02），纵轴为该权重区间内所包含西医症状个数占该区间内质量短语个数的比例，而折线图中点上的值为该区间下质量短语的个数。例如，点（0.94，0.83 6）表示在权重区间［0.94，0.96］中西医症状的比例为 0.83，同时该区间内共有 6 个质量短语。

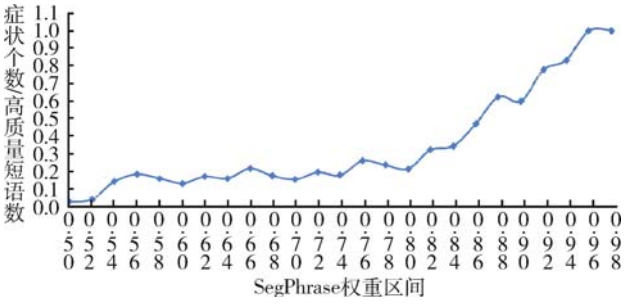

图 4 高质量短语列表中权重区间和西医症状比例的关系

由图 4 可知，当权重 w 取 0.86 时，西医症状比例就下降到了 0.5 以下，但此时能收集到的症状个数过少，不适合作为阈值；当 $w > 0.54$ 时，西医症状比例在 0 ~ 0.2 之间，过于稀疏。基于以上分析，本文将阈值设为 0.54，对大于该阈值的高质量短语进行简单过滤，最终得到 400 个西医症状实体。

4.3 基于模式的症状识别结果

为了验证基于症状构成模式的症状识别方法中每一步的有效性，本文设计了两项中间步骤的试验与本文方法进行对比。首先，只用从医疗网站上抽

取得到的医疗实体子集使用基于词典分词的方法识别病历文本中的症状；然后，把 SegPhrase 方法学习到的西医症状子集加入其中，即使用本文构建的西医症状词库识别症状实体；最后，加上基于模式的后处理进行试验。这组试验使用 4.1 节所说的 20 份测试数据，共包含 2 731 条句子、589 个西医症状实体。结果见表 3，由表 3 可知仅基于医疗实体子集识别症状，*F* 值最高只有 0.261 6，而加入 SegPhrase 方法收集的西医症状子集后，正确率（Precision）和召回率（Recall）都增加了将近 1 倍。虽然 SegPhrase 方法仅仅收集了 400 个西医症状实体，但该方法学习出的实体均是电子病历中高频率的西医症状实体，因此在症状实体识别过程中贡献了很多有效实体。最后将基于模式的后处理方法加入实验，基于 BDMaxMatch 方法的准确率达到了 0.947 6，对应 *F* 值为 0.884 1，这说明针对病历文本中构成复杂的症状实体，基于 Pattern 的方法可以有效识别出它们的边界，进一步提升了识别效果。对比 BDMaxMatch 和 Ansj 的识别结果，基于词典的机械分词方法要好于添加了用户自定义词典的分词工具，这说明分词工具可能适用于对新闻等普通文本的分词，而对病历文本这类特殊语料进行分词反而会引入歧义，减弱识别效果。

本文还使用训练集[12]训练出 CRF 模型，使用测试数据进行评测，结果如表 3 中最后一行所示。与本文方法相比，监督学习方法的 *F* 值比文本方法的最好结果低了 3%，整体效果也不如本文方法。

表 3 基于症状构成模式的电子病历症状识别方法的对比实验结果

项目	BDMaxMatch			Ansj		
	正确率	召回率	*F* 值	正确率	召回率	*F* 值
医疗实体子集	0.349 7	0.169 8	0.228 5	0.314 3	0.224 1	0.261 6
+SegPhrase 收集西医症状子集	0.603 4	0.534 8	0.567 1	0.564 2	0.514 4	0.538 2
+基于模式的后处理	0.947 6	0.828 5	0.884 1	0.879 7	0.794 6	0.835 0
CRF 模型	0.894 6	0.817 0	0.854 0	–	–	–

5 结语

本文提出了基于症状构成模式的方法自动识别中文电子病历文本中的症状实体，该方法具有良好的可扩展性并与监督学习方法对比证明了方法的有效性，下一步的工作将利用该方法进行中医症状实体的抽取。此外，基于本文在构建西医症状词库时得到的症状知识库，可以进一步扩充，形成一个相对完善的医疗领域知识库。

参考文献

1 World Health Organization. The ICD－10 Classification of Mental and Behavioral Disorders: diagnostic criteria for research [J]. Geneva World Health Organization, 1993, 8 (12): 14.

2 McDonald C J, Huff S M, Suico J G, et al. LOINC, a Universal Standard for Identifying Laboratory Observations: a 5－year update [J]. Clinical Chemistry, 2003, 49 (4): 624－633.

3 Darden T, York D, Pedersen L. Particle Mesh Ewald: an N·log (N) Method for Ewald Sums in Large Systems [J]. The Journal of Chemical Physics, 1993, 98 (12): 10089－10092.

4 Donnelly K. SNOMED－CT: the advanced terminology and coding system for eHealth [J]. Studies in Health Technology and Informatics, 2006, (121): 279.

5 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology [J]. Nucleic Acids Research, 2004, 32 (suppl 1): 267－270.

6 许华, 刘茂福, 姜丽, 等. 基于语言规则的病症菌实体抽取 [J]. 武汉大学学报: 理学版, 2015, 61 (2): 151－155.

7 Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora [C]. Beijing: 14th International Conference on Computational Linguistics, 2010.

8 Mintz M, Bills S, Snow R, et al. Distant Supervision for Relation Extraction without Labeled Data [C]. Singapore: Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, 2009.

9 Xu D, Zhang M, Zhao T, et al. Data - Driven Information Extraction from Chinese Electronic Medical Records [J]. Plos One, 2014, 10 (8): e0136270.

10 刘凯, 周雪忠, 于剑, 等. 基于条件随机场的中医临床病历命名实体抽取 [J]. 计算机工程, 2014, (9): 312 - 316.

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

12 Wang Y, Yu Z, Li C, et al. Supervised Methods for Symptom Name Recognition in Free - text Clinical Records of Traditional Chinese Medicine: an empirical study [J]. Journal of Biomedical Informatics, 2014, 47 (2): 91 - 104.

13 徐永东, 权光日, 王亚东. 基于 HL7 的电子病历关键信息抽取技术研究 [J]. 哈尔滨工业大学学报, 2011, 11: 89 - 94.

14 Métivier J P, Serrano L, Charnois T, et al. Automatic Symptom Extraction from Texts to Enhance Knowledge Discovery on Rare Diseases [M]. Berlin: Springer International Publishing, 2015: 249 - 254.

15 Okumura T, Tateisi Y. A Lightweight Approach for Extracting Disease - Symptom Relation with MetaMap toward Automated Generation of Disease Knowledge Base [M]. Berlin: Springer Berlin, 2012: 164 - 172.

16 Aronson A R. Automatic MeSH Term Assignment and Quality Assessment [C]. Washington, DC: Amia Symposium, 2001.

17 Aronson A R. Metamap: Mapping Text to the UMLS Meta Thesaurus [EB/OL]. [2016 - 02 - 02]. https://ii.nlm.nih.gov/Publications/Papers/metamap06.pdf.

18 Bing L, Chaudhari S, Wang R C, et al. Improving Distant Supervision for Information Extraction Using Label Propagation through Lists [C]. Lisbon: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015.

19 Bing L, Ling M, Wang R C, et al. Distant IE by Bootstrapping Using Lists and Document Structure [EB/OL]. [2016 - 01 - 10]. http://arxiv.org/pdf/1601.00620v1.pdf

20 Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge [C]. Vancouver: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008: 1247 - 1250.

21 Liu J, Shang J, Wang C, et al. Mining Quality Phrases from Massive Text Corpora [C]. Melbourne: ACM SIGMOD International Conference on Management of Data, 2015.

22 Blei D M. Dynamic and Supervised Topic Models for Literature - Based Discovery [R]. Princeton, New Jersey, 2011.

2016 年《医学信息学杂志》编辑出版重点选题计划

2016 年本刊将继续以“学术性、前瞻性、实践性”为特色，及时追踪并深入报道国内外医学信息学领域前沿热点，反映学科研究动态，展示学科应用成果，引领学科发展方向。现对 2016 年度编辑出版重点选题策划如下：

一、医药卫生体制改革与医药卫生信息化

1“十三五”卫生信息化建设的创新与发展；2 医药卫生信息规划与发展战略；3 区域卫生、公共卫生、基层卫生信息化建设；4 各级医疗健康信息平台建设；5 医疗卫生信息相关标准研发、应用和落地；6 医疗卫生信息化相关安全隐私保护和法律法规；7 国外医药卫生信息化建设最新技术、成功经验。

二、医学信息技术

1 基于健康大数据的科学决策与监管；2 医学大数据与精准医疗；3“互联网 +” 医疗；4 移动医疗、远程医疗服务与健康管理；5 物联网、智慧医疗技术与实现；5 各类医学信息系统信息互通与操作衔接；6 医学机构知识库构建技术与方法。

三、医学信息研究

1 医学信息学理论及方法研究；2 医学科技创新体系和发展战略；3 医学科技监测与舆情监测；4 医药卫生信息分析评价；5 生物医学数据挖掘与利用、知识发现技术与实现。

四、医学信息组织与利用

1 医学数字图书馆发展趋势与标准建设；2 泛在化医学知识服务与决策咨询服务；3 医学知识组织的关键技术与发展方向；4 医学信息交互及存取；5 医学图书馆区域合作及资源共享模式研究。

五、医学信息教育

1 医学信息专科、本科、研究生教育及继续教育体制改革与模式创新；2 医学信息素养及职业岗位的培养与教育；3 医学信息课程改革与实践；4 国外医学信息学教育的先进经验借鉴。

（《医学信息学杂志》编辑部）