·信息研究·

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

许　源[1]，葛艳秋[2]，王　强[3]，熊　刚[1]，易应萍[1,4]
(1. 南昌大学附属第二医院临床医疗大数据研究中心，江西 南昌 330006；2. 南昌大学医学部公共卫生学院，江西 南昌 330006；3. 赫博特医疗信息科技有限公司，江苏 苏州 215000；4. 南昌大学第二附属医院科教处，江西 南昌 330006)

摘　要：【目的】研究针对非结构化临床电子病历的自然语言处理模型的构建和优化，并利用该模型对江西省医疗大数据平台中卒中病人的病历进行结构化数据提取。【方法】从江西省医疗大数据平台中随机筛选 500 份 2011-2016 年的卒中病人入院记录，根据临床科研的实际需求构建了脑卒中专科病人的命名实体标注体系和命名实体标注语料库，利用该语料库构建基于 CRF 以及 RUTA 规则的命名实体抽取模型，并通过调整 RUTA 规则以及参数提升识别准确率。【结果】经五折交叉验证，该模型的医学命名实体的抽取准确率 0.960，召回率 0.916，*F*-score 0.939，利用该抽取模型对大数据平台中 10 295 份脑卒中患者入院记录进行抽取，共抽取命名实体 264 580 条，命名实体修饰 1 161 077 条。【结论】构建的自然语言抽取模型识别准确率较高，通过该模型能够准确地从大量非结构化病历中获取病人的既往史、生活史、临床表现等有价值的科研数据，有效提升心脑血管疾病的临床科研效率和科研水平。

关键词：中文电子病历；命名实体识别；条件随机场 CRF；脑卒中
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1672-3554(2018)03-0455-08

Medical Name Entity Recognition and Application in Chinese Admission Record of Stroke Patients Based on CRF and RUTA rule

XU Yuan[1], GE Yan-Qiu[2], WANG Qiang[3], XIONG Gang[1], YI Ying-Ping[1,4]
(1.Clinical Big Data Research Center, The Second Affiliated Hospital, Nanchang University, Nanchang 330006, China; 2.Department of Public Health, School of Medicine, Nanchang University, Nanchang 330006, China; 3.HBT Medical Information Company, Suzhou 215000, China; 4.Department of Science and Education, The Second Affiliated Hospital of Nanchang University, Nanchang 330006, China)
Corresponding to: YI Ying-Ping; E-mail: yyp66@126.com

Abstract: 【Objective】To research the construction and optimization of natural language processing model for unstructured medical records, and using the model to extract structured data from medical records of stroke patients in Jiangxi Medical Big Data Platform. 【Methods】According to the actual needs of clinical research, a stroke specialist entity annotation system and named entity annotation corpus were constructed based on 500 hospital admission records of stroke patients, which randomly selected between 2011 to 2016 from the Jiangxi provincial medical big data platform. The corpus is used to construct a named entity extraction model based on CRF and RUTA rules, and the recognition accuracy is improved by adjusting RUTA rules and parameters. 【Results】Accuracy rate of extraction model was 0.960, recall rate was 0.916 and F-score was 0.939. The extraction model was used to extract 264 580 entities and 1 161 077 entity relation

收稿日期：2017-12-15
基金项目：江西省科技厅科技创新平台(20171BCD40024)；江西省科技厅一般项目(20171BBH80025)
作者简介：许源，硕士生，研究方向：医学信息、人工智能，E-mail：13810308036@126.com；易应萍，通信作者，研究员，研究方向：医疗大数据、数据挖掘、医院管理，E-mail：yyp66@126.com

from 10 295 stroke patients' admission records of the medical big data platform.【Conclusions】The constructed natural language extraction model has a high recognition accuracy, which can accurately obtain valuable scientific research data of patients' past history, life history and clinical manifestations from a large number of unstructured medical records and effectively improve the clinical research efficiency and scientific research level of cerebrovascular diseases.

Key words: Chinese medical record; named entity recognition; CRF; stroke

[J SUN Yat-sen Univ(Med Sci),2018,39(3):455-462]

电子病历是临床医生在治疗过程中产生的医疗记录,海量电子病历数据包含了大量的医疗知识和患者的健康信息应得到有效利用。目前自然语言处理技术已经广泛应用于英文电子病历的临床医学研究中,并且已有统一的病历书写标准UMLS以及I2B2等组织发布的公开医疗病历语料库[1],而中文的公开医疗语料库几乎没有,因此针对中文电子病历的自然语言识别的相关研究和文章也较少[2]。目前主流的中文病历识别方式是条件随机场(CRF)[3-4],该方法已被广泛应用于英文医学命名实体的识别[5-6]。除了单纯的改进标注模式优化模型参数之外,很多研究机构还探索出各种利用CRF的混合优化方案[7],如冯丽芝[8]利用CRF+Bootstrapping的混合命名实体识别方法;栗伟等[9]利用基于CRF与规则相结合的医学病历实体识别算法,采用CRF进行病历实体的初始识别,然后基于规则进行病历实体识别结果优化。程健一等[10]也提出了一种基于支持向量机(SVMs)和条件随机场(CRF)双层分类模型的去隐私化方法。2013年后,中文电子病历命名实体识别的研究命名算法包括SSVM、CRF及其加rule或bootstraping或双解构法、HER神经网络等,准确率为0.696~0.935之间[9-12]。目前针对中文临床文本数据的NLP技术,包括:"医学命名实体识别","命名实体编码","命名实体修饰词识别","时间信息抽取"等[13],已积累了一些技术知识储备,但是目前中文自然语言抽取模型还没有被广泛应用到临床科研中。因此,本研究希望构建单病种(卒中)中文临床病历语料库,以丰富相关临床研究数据资源。

1 材料与方法

1.1 语料数据来源

南昌大学第二附属医院是江西省集医疗、教学、科研为一体的综合三级甲等医院,年住院10万余人次。随着电子病历的普及,每年产生的电子病历数据快速增长,2011-2015年产生电子病历数量分别为21、59、86、117、250万份。本项研究数据均来自南昌大学第二附属医院医疗大数据工程技术研究中心,由于本研究仅针对一种疾病类型构建自然语言抽取模型,因此500份入院记录能够涵盖所有会出现的症状、用药、病史的情况。采用随机抽样的方式从2011-2015年间收治的脑卒中患者中抽取500份入院记录作为标注语料,筛选条件为患者出院诊断的诊断编码中ICD10类目码为I61、I62、I63的病人,所有数据均进行了脱敏处理。

1.2 PLATO医学自然语言处理系统

PLATO系统是苏州赫博特医疗信息科技有限公司自主研发的自然语言处理系统,该系统集成了文本标注、交叉验证、测试集验证等模块,同时平台可集成机器学习、深度学习算法快速训练自然语言抽取模型并对模型准确率进行检验。通过该平台将非结构化的医疗文本被转化为包含重要医学信息的结构化数据后可以xml格式输出或存储于数据库中。

1.3 命名实体及命名实体修饰分类设计

目前自然语言处理技术主要分为有监督学习和无监督学习两类[14]。本项研究的目的是通过自然语言处理技术获取电子病历中的病人信息,提供更加全面的科研数据支持,处理对象是脑卒中病人的入院记录,类型较为单一,且存在较多的专业术语,因此用监督的机器学习方式准确性更高、更适合于本项研究[15]。对于有监督的机器学习方式,制定合理的标注体系和较高的标注一致性才能保证机器识别的准确性,因此本项目邀请了多名神经内科的专家共同参与标注体系的构建与研究,包括两位神经内科主任医生与三名神经内科博士,经过多次研究讨论,并结合我院临床科研的

实际需求将入院记录中的命名实体分为5类，同时定义了5类命名实体修饰分类。

1.3.1　命名实体分类　本研究借鉴UMLS对概念类型的定义方法，将入院记录中的命名实体分为5类，包括危险因素、临床症状、检验检查、治疗、药物[16]。该命名实体分类基本涵盖了入院记录中大部分有价值的信息，同时将药物单独分类有利于后期临床研究的推进，符合医疗领域的书写惯例，有很好的可操作性。本研究所定义的命名实体遵循实体间不重叠、不嵌套、实体内不含有表示停顿的标点符号（比如逗号、句号、顿号等）这3个原则。①危险因素（Risk）：危险因素包括所有已知的、可能的或待发现的导致的脑卒中的致病因素，如高血压（血压控制不佳）、高血糖、高血脂、抽烟、饮酒、家族遗传、脑动脉硬化（狭窄）、颈动脉狭窄、既往脑梗、冠心病（支架）、风心病、心脏病、偏头痛、颅底血管闭塞症（烟雾病）、房颤、吸毒、同型半胱氨酸、慢性支气管炎、哮喘、颅脑外伤、环境物理或化学伤害、伴既往心脑血管病、绝经、流产等。②临床症状及表现（Symptom）：所有病人相关的临床表现，如头晕、头痛、恶心、呕吐、咳嗽、咳痰、寒战、抽搐、发热、腹痛、口吐白沫、流涎、乏力、麻木、口角歪斜、舌偏、口齿不清、听力下降、大小便失禁、突发记忆力下降、认知障碍、焦虑、抑郁、烦躁不安、神志不清、精神差、睡眠不佳、食欲不振等。③检验检查（Labtest）：包括临床检验和医技检查，如CT、MRI等。④手术（Procedure）：患者入院前所采取的相关手术治疗。⑤药物（Medicine）：患者入院前所服用相关的药物，如“口服硝苯地平缓释片降压”中的“硝苯地平缓释片”。

1.3.2　命名实体修饰分类　本研究对命名实体修饰也进行了定义，命名实体修饰反映了命名实体与患者的具体关系，如是否患病、发病部位、发病时间等等[15]。因此本研究将命名实体修饰分五类，包括否定前缀、部位、时间点、时间段以及数值。①否定（Negation）：由于标准化的入院记录模板中有大量否定描述，因此将否定前缀等，如“否认”、“无”等均属于否定前缀。②部位（Body-parts）：表明症状的出现部位及危险因素中某些疾病的发病部位，如“左侧”、“双侧”、“脑室旁”等。③时间点（Time-point）：症状或危险因素发生的时间点，如某年某月、某年秋天等。④时间段（Time-period）：表示症状持续的时间，以判断患者是否在窗口期进行了有效治疗。在标注时应选择如：3 h（3 h前）、3 d（3 d前）、3年（3年前）之类的词段。⑤数值（Value）：数值包括病情检查的数值、药物用量的数值、血压数值等。本研究对命名实体与修饰间的关系进行了限定，确保标注人员以及模型中不存在无关的修饰，修饰关系如图1所示。

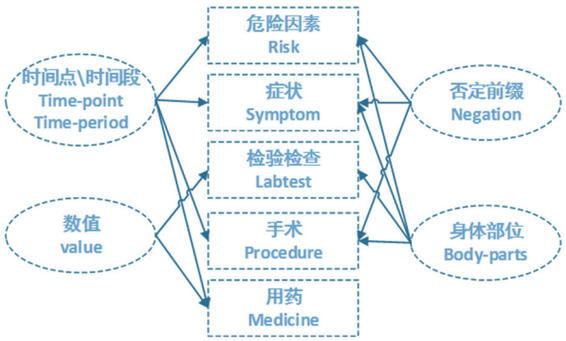


图1　命名实体与修饰间的关系限定

Fig.1　Relationship restriction between the named entity and the modifier

1.4　语料标注

语料标注流程语料标注工作是在两名有医学背景的硕士研究生的帮助下完成的，我们制定了完善的命名实体标注规范流程，并对两名标注人员进行标注体系和标注系统的培训，保证标注的一致性和准确性。具体流程如下图所示，首先从500份脑卒中病人入院记录中随机抽取150份，每人标注100份入院记录，其中50份相同，用于一致性验证。当标注的一致性达到0.9时才进行剩余的入院记录标注，否则重新学习标注规则，再进行标注，标注流程如图2所示。

对于标注的语料中与临床科研无关、症状正常以及描述不准确的内容不进行标注，如“讲话语速慢”、“神志清楚”、“大便如前”、“某院住院”、“初步判断”等，既提高了标注的效率，也减少了垃圾数据的产生。本次标注共耗时9 d，标注结果如图3A所示，每份病历大致标注200个左右命名实体和100个左右的命名实体修饰，如图3B所示。

1.5　语料标注一致性评价

本项研究采用几个标准的系统评测指标对标注结果进行评估，包括准确率（$P=A_3/A_2$）、召回率（$R=A_3/A_1$）以及F值（$F=2P\cdot R/(P+R)$）。对两名标

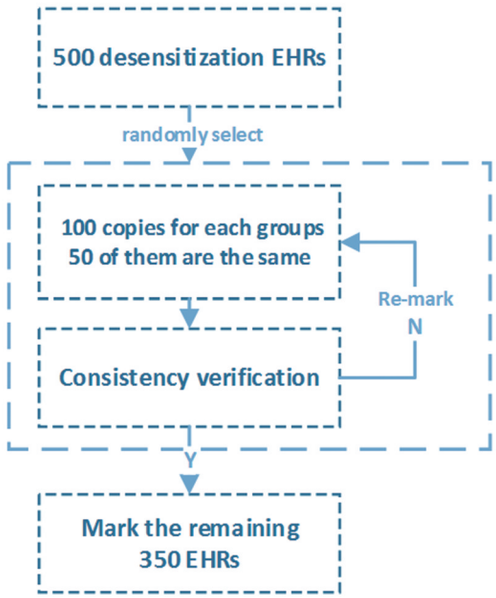


图2 标注流程

Fig.2 Mark process

注人员标注相同的50份入院记录进行评估,将其中一人标注的命名实体作为标准数量(A_1),另一人标注的命名实体数量作为预测数量(A_2)以及两人标注一致的数量(A_3),计算相应的准确率、召回率、F值[17],当标注一致性F值大于0.9时我们才对剩余的语料文档进行标注。

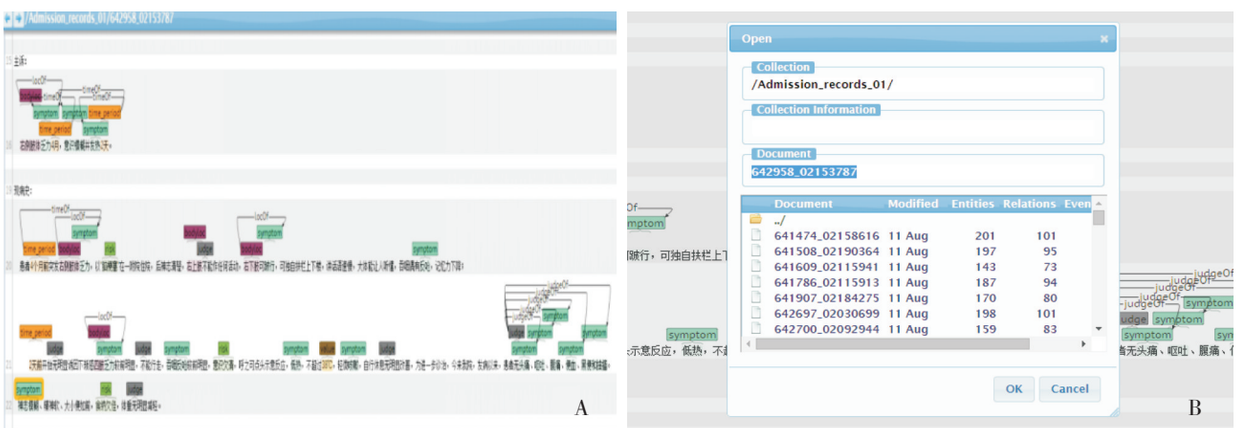
1.6 基于CRF构建自然语言处理模型

在处理标注类问题时候,通常把此类问题称之为"结构化预测",该类问题与其他标注的分类问题有着很大的不同。其中包括:①词和词之间并非是独立的,标准分类问题一般是对每一个词分别预测,看成是一个一个的独立词,会丢失许多信息。②预测时,句子的长度不同,不方便将句子统一成相同的向量。③标注序列的解集与句子长度呈指数增长,穷举法计算无法使用。通过条件随机场(CRF)就可以很好地解决上述存在的问题。CRF可被定义为一个无向图模型,最早由Lafferty和McCallum在提出,是由一个点序列所表示的随机变量Y、由边表示的条件依赖关系以及任意序列的条件X所组成,随机变量只与其相邻元素有条件依赖关系(图4)。因此,条件随机场本质上是一个在给定输入节点序列条件下计算输出节点条件概率的无向图网络,而命名实体的识别就是一个序列标注问题,因此本次研究采用线性链条件随机场。

在给定输入序列X后,标记序列Y的条件概率可被定义为下式:

$$P(y \mid x,w)=\frac{\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)}{Z(x,w)} \quad (1)$$

其中,规范化因子$Z(x,w)=\sum_i \exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$,$t_k$和$s_l$都是特征函数,$\lambda_k$和$\mu_l$是对应的权值参数。对于CRF模型,通常有3个问题,第一个问题是关于通过给定输入序列X,标注序列Y,以及特征函数$F(x,y)$来求解模型的参数w,这个被称为CRF的学习问题,一般是采用改进的迭代尺度法(IIS)来求解出最优的模型参数。第二个问题是关于给定模型参数w,



A: Name entity mark sample; B: mark result (include the quantity of entities and relations)

图3 入院记录文档标注图例

Fig.3 Hospital admission record marked legend

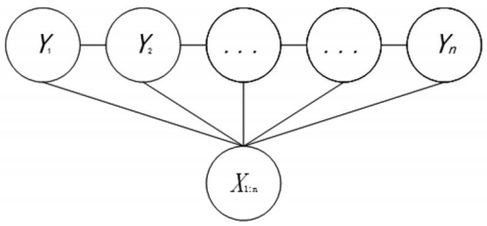


图4　CRF无向图的链式结构

Fig.4　CRF chain structure of undirected graph

以及输入序列X，求解出概率最大的标注序列Y，这个问题被称为CRF的预测问题，一般采用维特比算法来求解出最优的输出序列y。第三个问题是对于给定模型参数w，以及输入序列X，求解出特定标注序列Y的条件概率，这个被称为CRF的概率计算问题，通常使用前向-后向算法来求解出特定序列Y的概率。

在CRF模型训练时需调整参数来提高模型的抽取准确率，由于本研究脑卒中入院记录语料库的语料数据量较大，因此对参数w采用L-BFGS算法进行求解，该算法在计算的时候既能够收敛得比较快，在存储的时候又可以节省大量的空间。另一训练参数C初始设定为0.9，用来防止训练过程的过拟合现象。通过交叉验证（Cross-validation）的方法来对模型的泛化能力进行测试，由于数据量较大，选择5-fold的方法进行交叉验证，最终将提取到五组的准确率的均值作为模型最终的准确率。

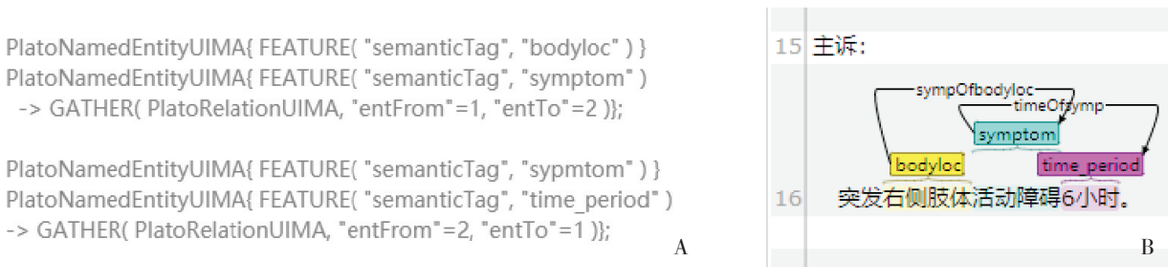
1.7　结合RUTA规则

在医学命名实体识别领域，条件随机场（CRF）是应用最为广泛的有监督机器学习算法，许多研究者也在探索CRF与各类其他算法相结合的混合优化方案[18]。针对本研究的脑卒中入院记录语料库建立基于CRF的自然语言抽取模型，通过添加各种规则辅助CRF预测模型，以此提高提取模型的准确率。目前可在抽取模型中增加的规则包括：字典匹配，否定含义分析，以及RUTA规则。其中的RUTA规则功能尤为强大，RUTA规则是一种基于角色的脚本语言，其中包含正则表达式以及自身所构建的语法。RUTA规则不仅可以对特定信息进行标注，也可以对特定关系进行给定。在病历自然语言提取模型中，通过分析病历的特点，编写特定的规则来增加和删除符合特定条件的命名实体，更重要的是可以利用RUTA规则里关系的规则。例如主诉“突发右侧肢体活动障碍6小时”这句话，其中包含了症状，疾病部位，持续时间这3个实体信息，仅仅通过CRF模型，并不能将整个句子的完整含义提取出来，只能独自的实体名称，因此会丢失很多重要的信息，引入了RUTA规则以后，通过编写特定的RUTA关系规则，能够将实体之间的关系一起提取出来，如图5所示是具体的RUTA规则以及实现的效果，通过匹配连续的“bodyloc”和“symptom”标签，以及连续的“symptom”和“time_period”的标签，将具体的部位和症状以及时间和症状等包含了前后关系的命名实体一同提取出来，因此可以大大提高各个实体之间的连通性，并且减少了提信息的损失。

2　结　果

2.1　语料标注结果

对初期两名标注人员所标注完成的50份相同的入院记录文本进行命名实体和命名实体修饰



A：RUTA code sample；B：CRF+RUTA executed result

图5　RUTA规则示例图

Fig.5　RUTA rules sample diagram

进行一致性验证。一致性验证的结果中标注准确率为0.929、召回率为0.884、*F*值为0.906达到进行下一步标注的一致性需求(表1)。

2.2 模型识别准确率

当在PLAOT中仅仅只采用CRF模型进行命名实体识别时,对比不同训练样本数生成的模型,可以发现当电子病历数量样本为50份时,准确率仅为0.839,因此在建立提取命名实体的模型时,通过追加字典和RUTA规则以及样本数量,模型识别的准确率经过五折交叉验证准确率提升到了0.939,对比的结果如表1所示。详细的各命名实体和命名实体修饰的识别准确率如表2所示。

表1 不同样本数的训练模型优化前后的F值对比

Table 1 Comparison of F values before and after training model optimizationwith different numbers of samples

Quantity	CRF	CRF+rule	CRF+RUTA
50	0.839	0.856	0.906
100	0.853	0.865	0.907
150	0.869	0.897	0.903
200	0.876	0.897	0.910
500	0.893	0.902	0.939

表2 500份入院记录命名实体及名实修饰抽取的交叉验证准确率

Table 2 Name entitiesand modifiersextraction cross-validation accuracy of500 hospital admission record

No.	Entity type	Quantity	Accuracy	Recall	F score
1	症状(Symptom)	24 665	0.97	0.97	0.97
2	手术(Procedure)	432	0.90	0.75	0.82
3	检验检查(Labtest)	9 170	0.97	0.94	0.96
4	危险因素(Risk)	15 042	0.93	0.91	0.92
5	药物(Medicine)	1 008	0.95	0.800	0.87
	Total name entity	50 317	0.96	0.92	0.93
6	否定词(Negation)	27 462	1.00	1.00	1.00
7	时间点(Time_point)	3 407	1.000	0.95	0.97
8	身体部(Bodyloc)	20 620	0.97	0.96	0.96
9	数值(Value)	5 146	0.98	0.96	0.97
10	时间段(Time_period)	1 430	0.96	0.94	0.95
	Total modifiers	58 065	0.96	0.93	0.96
	Totals	108 382	0.96	0.92	0.94

2.3 模型应用

利用基于CRF+RUTA规则的混合自然语言处理模型对7 369名脑卒中患者的10 295条入院记录进行相关信息进行提取,获取命名实体264 580条、命名实体修饰1 161 077条,大量丰富了临床科研医生的科研数据总量(表3)。

表3 10295份脑卒中入院记录命名实体及命名实体修饰抽取统计表

Table 3 Named entities and modifiersextract statistics of 10295 Stroke hospitaladmission records

Type	Item	Quantity
Name Entity	症状(symptom)	77 635
	风险(risk)	156 567
	药物(medicine)	21 740
	检验\手术(labtest\procedure)	8 638
Total		264 580
Name entity modifier	否定词+风险(negOFrisk)	170 736
	否定词+检测(negOflabtest)	49 172
	否定词+症状(negOfsymptom)	439 085
	检测+风险(labtestOfrisk)	25 135
	身体部位+症状(bodyOfsymptom)	308 884
	时间+检测(timeOflabest)	6 184
	时间+手术(timeOfprocedure)	1 357
	时间+风险(timeOfrisk)	31 171
	时间+症状(timeOfsymptom)	27 932
	检测值+检测(valueOflabest)	97 572
	风险值+风险(valueOfrisk)	3 849
Total		1 161 077

3 讨 论

3.1 研究成果

在本次研究中,我们与多位神经内科专家共同探讨了脑卒中临床科研的相关需求,同时参考2010年I2B2挑战的标注模式,构建了一套适合用于临床科研的快速标注体系。但不同于2010年I2B2挑战的标注模式,我们并未对全部断言进行标注,如“可能”、“假设”、“目前”等状态词汇我们不进行标注,只对明确否定修饰进行标注,这部分数据存在主观判断,对于临床科研的意义不大,且有可能会影响最终抽取数据的准确性,因此我们只对病历中的有价值信息进行总结分类,最终对5类命名实体和5类命名实体修饰进行标注。通过制定合理的标注方式,两名标注人员对50份相

同的入院记录进行标注，标注准确率达到0.929，*F*值达到0.906，达到多人标注的一致性要求，两名标注人员耗时一周完成了剩余450份入院记录的命名实体和命名实体修饰的标记。针对已500份标注完成的入院记录，通过增加RUTA规则以及调整CRF参数，最终模型总体平均识别准确率为0.961、平均召回率0.916、平均*F* score 0.939。

在之前的研究中，已有研究人员利用SSVM也取得了不错的抽取准确率，但是SVM本质上仍是一个二元分类器，在面对多分类、巨量数据时，其低速和非平均分布数据的问题一直无法解决，由于本研究构建的抽取模型最终将在大数据平台中处理海量的电子病历文本，因此本研究采用CRF算法作为自然语言模型的核心算法。通过加入了特定字典以及RUTA规则等，不断的补充提取模型，最后的准确率高于其他单纯使用CRF算法进行的电子病历自然语言抽取研究，甚至高于部分混合优化算法的研究。原因有下几点：首先，相较于其他研究对各个科室所有电子病历随机进行抽取的方式，我们对标注的原始病历语料进行了筛选，只抽取诊断包含脑卒中相关疾病的病人，使得出现未标注专业术语的概率降低；其次，我们制定了适合临床科研的标注体系，对于一些不明确、没有科研价值的语句不进行标注，减少了标注的种类；最后，我们进行了严谨的一致性验证，让两名标注人员分别对50份相同的病历进行标注，当两名标注人员的标注一致性达到要求才进行下一步研究。

3.2　局限性和展望

本项目的研究目标与I2B2挑战和一些其它测评机构的研究目的不尽相同，I2B2的命名实体识别评测挑战中仅对医疗问题、检验和治疗三个概念类进行标注和识别，命名实体修饰也仅限于医疗问题的修饰，目的是探索优化算法和标注模式以提高模型识别准确率[19-20]，而本项目旨在以最有效且识别率最高的方式获取入院记录中与脑卒中研究相关的临床信息[21]，并将该自然语言识别模型应用于大数据研究平台，为江西省未来脑卒中相关研究提供更为丰富的数据来源。

我们目前只针对脑卒中病人的病历数据进行了标注，构建了单病种的自然语言抽取模型，该模型对于其它疾病的入院记录识别准确率可能会降低。未来我们将会把该方法扩展到其他类型的文本数据，如磁共振报告、CT报告、手术记录等，使临床研究工作人员能够在科学研究的过程中更加有效的获取有价值的医疗信息。

3.3　结　论

我们根据神经内科医生的实际科研需求，建立了脑卒中患者入院记录快速标注体系和标注流程，对500份脑卒中病历进行标注，标注*F*值0.906。基于已标注好的语料库，结合条件随机场算法，构建脑卒中专科的自然语言抽取模型，获得总体平均准确率0.961、平均召回率0.916、平均*F* score 0.939。最后利用该抽取模型对江西省大数据平台中的10 295份脑卒中病历进行自然语言抽取，共获取命名实体和命名实体修饰数据116万条，极大地丰富了脑卒中相关临床研究的数据资源。

参考文献

[1] 简哲，李燕. 电子病历自然语言处理测评发展[J]. 医学信息学杂志，2016，37(12)：10-13.
Jian Z，Li Y. Development and evaluation of natural language processing of electronic medical record [J]. J Med Inform，2016，37(12)：10-13

[2] 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[J]. 自动化学报，2014，40(8)：1537-1562.
Yang JF，Yu QB，Guan Y，et al. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction [J]. Acta Automatica Sinica，2014，40(8)：1537-1562.

[3] 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，2014，40(9)：312-316.
Liu K，Zhou XZ，Yu J，et al. Named entity extraction of traditional Chinese medicine medical records based on conditional random field [J]. Comp Engin，2014，40(9)：312-316.

[4] 张坤丽，马鸿超，赵悦淑，等. 基于自然语言处理的中文产科电子病历研究[J]. 郑州大学学报(理学版)，2017，49(4)：40-45.

Zhang KL, Ma HC, Zhao YS, et al. Study on Chinese obstetric electronic medical records based on natural language processing [J]. J Zhengzhou Univ (Sci Ed), 2017, 49(4): 40-45.

[5] Urbain J. Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models [J]. J Biomed Inform, 2015, 58 Suppl(S):S143-S149.

[6] Tseytlin E, Mitchell K, Legowski E, et al. NOBLE- Flexible concept recognition for large-scale biomedical natural language processing [J]. BMC Bioinformatics, 2016,17(1):32-40.

[7] 燕杨, 文敦伟, 王云吉, 等. 基于层叠条件随机场的中文病历命名实体识别[J]. 吉林大学学报(工学版), 2014, 44(6):1843-1848.
Yan Y, Wen DW, Wang YJ, et al. Named entity recognition in Chinese medical records based on cascaded conditional random fields [J]. J Jilin Univ (Engin Technol Ed), 2014, 44 (6): 1843-1848.

[8] 冯丽芝. 面向命名实体抽取的大规模中医临床病历语料库构建方法研究[D]. 北京交通大学, 2015.
Feng LZ. Automatic approaches to develop Large-scale TCM electronic medical records corpus for named entity recognition tasks [D]. Beijing Jiaotong University, 2015.

[9] 栗伟, 赵大哲, 李博, 等. CRF与规则相结合的医学病历实体识别[J]. 计算机应用研究, 2015, 32(4):1082-1086.
Li W, Zhao DZ, Li B, et al. Combining CRF and rule based medical named entity recognition [J]. Appl Res Comp, 2015, 32(4): 1082-1086.

[10] 程健一, 关毅, 何彬. 基于SVM和CRF双层分类器的英文电子病历去隐私化[J]. 智能计算机与应用, 2016, 6(6):17-19.
Cheng JY, Guan Y, He B. De- identification on electronic medical records using a two tier classifier based on SVM and CRF [J]. Intell Comp Appl, 2016, 6(6):17-19.

[11] Wang H, Zhang W, Zeng Q, et al. Extracting important information from Chinese operation notes with natural language processing methods [J]. J Biomed Inform, 2014,48(C):130-136.

[12] Zhao C, Jiang J, Xu Z, et al. A study of EMR-based medical knowledge network and its applications [J]. Comput Methods Programs Biomed, 2017,143(C):13-23.

[13] He B, Dong B, Guan Y, et al. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts[J]. J Biomed Inform, 2017,69(C): 203-217.

[14] 奚雪峰, 周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报, 2016,42(10):1445-1465.
Xi XF, Zhou GD. A survey on deep learning for natural language processing [J] .Acta Automatica Sinica, 2016, 42(10): 1445-1465

[15] 吴嘉伟. 电子病历实体关系抽取研究[D]. 哈尔滨工业大学, 2014.
Wu JW. Electronic medical records entity relationship extraction [D]. Harbin Inst Tech, 2014.

[16] 蒋志鹏,赵芳芳,关毅,等.面向中文电子病历的词法语料标注研究[J]. 高技术通讯, 2014,24(6):609-615.
Jiang ZP, Zhao FF, Guan Y, et al. Learning lexical corpus for Chinese electronic medical record [J].High Tech Letters, 2014, 24(6): 609-615.

[17] 马敬东, 梁力凡, 夏晨曦. 电子病历自由文本实体关系抽取[J]. 医学信息学杂志, 2016, 37(12):2-9.
Ma JD, Liang LF, Xia CX. Entity relation extraction for free text in electronic medical records [J]. J Med Intellig, 2016, 37(12): 2-9.

[18] 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.
Yang JF, Guan Y, He B, et al. Corpus construction for named entities and entity relations on Chinese electronic medical records[J]. Ruan Jian Xue Bao, 2016,27(11): 2725-2746

[19] de Bruijn B, Cherry C, Kiritchenko S, et al. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010[J]. J Am Med Inform Assoc, 2011, 18(5): 557-562.

[20] Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text[J]. J Am Med Inform Assoc, 2010,17(5):514-518.

[21] Castro VM, Dligach D, Finan S, et al. Large-scale identification of patients with cerebral aneurysms using natural language processing [J]. Neurology, 2017,88(2):164-168.

(编辑 刘清海)