doi:10.3969/j.issn.1672-1292.2022.01.012

基于 CRF 和深度学习的病历实体识别的研究

杨荣根[1],王　博[2],龚乐君[2]

(1.金陵科技学院智能科学与控制工程学院,江苏 南京 211169)
(2.南京邮电大学江苏省大数据安全与智能处理重点实验室,江苏 南京 210023)

[摘要]　随着电子病历数据量的快速增长,如何深层次、高效率地利用电子病历资源成为越来越迫切需要解决的问题. 从真实病历出发,研究电子病历的医学实体识别问题,为计算机更好地辅助医疗奠定基础. 通过人工标注的 108 份心血管科的真实病历数据与 3 类特征模板,运用条件随机场和双向长短时记忆网络联合条件随机场对心血管科电子病历疾病命名实体抽取的实验,并进行比较分析. 结果表明,结合合适的特征模版,条件随机场模型有更好的抽取性能,是一种较为适用的病历命名实体抽取方法.

[关键词]　电子病历,命名实体抽取,条件随机场,特征模板,双向长短时记忆网络
[中图分类号]TP391　[文献标志码]A　[文章编号]1672-1292(2022)01-0081-05

Research on Medical Record Entity Recognition Based on CRF and Bi-LSTM-CRF

Yang Ronggen[1],Wang Bo[2],Gong Lejun[2]

(1.College of Intelligent Science and Control Engineering,Jinling Institute of Technology,Nanjing 211169,China)
(2.Big Data Security and Intelligent Processing Key Laboratory of Jiangsu Province,
Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract:With the rapid increase in the amount of electronic medical record data,how to use electronic medical record resources in depth and efficiency has become more and more important. This article starts from the real medical record, through the manual annotation of 108 medical records of real medical records and three types of feature templates,using conditional random field and the bidirectional long-term short-term memory network conditional random field. Experiments on the extraction of cardiovascular electronic disease named entities and comparative analysis are conducted. The results show that CRF has better extraction performance,and that it is a more suitable method for extracting medical record named entities for small-scale and partially formatted medical record texts.

Key words:electronic medical record,named entity extraction,conditional random field,feature template,bidirectional long-term short-term memory network

电子病历(electronic medical record,EMR)被称为计算机化的病案系统或基于计算机的病人记录(computer-based patient record,CPR). 它是用电子设备(计算机、健康卡等)保存、管理、传输和重现的数字化的病人医疗记录,取代手写纸张病历. 它的内容包括纸张病历的所有信息. 美国国立医学研究所将其定义为:EMR 是基于一个特定系统的电子化病人记录,该系统提供用户访问完整准确的数据、警示、提示和临床决策支持系统的能力. 相较于传统的纸质病历,电子病历具有内容全面充分、书写标准规范、检索使用便利、存储更加简易,以及辅助临床诊断治疗等诸多优势. 同时电子病历对远程医疗有着积极的促进作用,电子病历可以为远程医疗提供快速、便捷、准确的病人资料,实现关键医疗信息的共享等优势. 因此,电子病历已经成为医疗卫生业的发展趋势,同时也成为了医院信息化的核心[1-2].

对医学命名实体的识别是实体关系抽取等任务的前提[3]. 医学命名实体识别的研究方法包括基于规则和统计的方法[4]、基于机器学习的方法,以及不依赖手工设计特征的深度学习方法[5]. 其中,李莹[6]利用条件随机场模型实现了基于语法语义的分析方式,从而实现了对医学问题和医学实体的识别和抽取. 栗伟[7]通过

收稿日期:2021-08-31.
通讯作者:杨荣根,博士,副教授,研究方向:学习理论、人工智能信息处理. E-mail:rg4592@jit.edu.cn

将基于 CRF 的实体抽取模型,并基于规则进行病历实体识别结果的优化. Lei 等[8]研究支持向量机(support vector machine,SVM)、最大熵(maximum entropy,ME)模型、条件随机场(conditional random field,CRF)、结构化支持向量机(structured SVM,SSVM)等机器学习方法,组合字袋、词袋、词性等不同的特征进行组合特征抽取电子病历中的医疗问题、医疗过程、用药等实体. 张晓斌[9]提出了一种卷积神经网络(convolution neural network,CNN)和长短期记忆(long short-term memory,LSTM)网络结合的实体关系抽取方法,在实体抽取上取得了较高的 $F1$ 值. 杨红梅[10]使用双向长短时记忆(bidirectional LSTM,Bi-LSTM)网络对肝癌病历进行的医学实体抽取. 曹春萍[11]提出一种集成卷积核神经网络(ensemble convolution neural network,ECNN)模型与双向长短期记忆网络和条件随机场结合的模型,该模型对临床病历文本中复合实体识别有较好的效果.

病历实体类型多样,数量众多,不断有未登录新词出现,因此 ICD-10 标准疾病库也不能涵盖所有疾病名称. 在疾病名称中,存在大量的嵌套、别名、缩略词等现象,没有严格可循的构词规律[12]. 这使得病历文本的抽取成为一项挑战性的工作,吸引了许多研究者投入这一领域. 另一方面,由于病历文本的隐私性以及专业性,尤其是真实病历获取十分困难,这也是病历实体抽取的难点之一. 本文从冠心病真实门诊病历出发,运用 CRF 及 Bi-LSTM-CRF 模型,对冠心病病历中的疾病实体进行识别和抽取.

1　模型

1.1　CRF 模型

条件随机场是给定随机变量 X 条件下,随机变量 Y 的马尔科夫随机场. 而其中用于序列标注的主要是线性链条件随机场. 如图 1 所示.

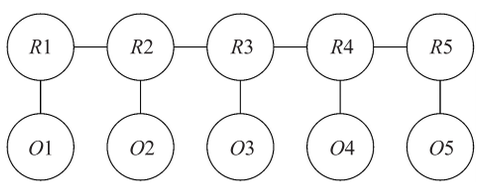

图 1　CRF 概率模型图

Fig. 1　CRF probability model diagram

设输入的观察序列 $O=\{O_1,O_2,\cdots,O_n\}$,预测序列 $R=\{R_1,R_2,\cdots,R_n\}$,

$$P(R_i\mid O,R_1,\cdots,R_{i-1},R_{i+1},\cdots,R_n). \tag{1}$$

当 $i=1,2,\cdots,n$ (在 $i=1$ 和 n 时只考虑单边)时,$P(R\mid O)$ 为线性链条件随机场. 当观察变量 O 取值为 o 的条件下,预测变量 R 取值为 r 的条件概率具有如下形式:

$$P(r\mid o)=\frac{1}{z(o)}\exp\Big(\sum_{i,k}\lambda_k t_k(r_{i-1},r_i,o,i)+\sum_{i,l}\mu_l s_l(r_i,o,i)\Big). \tag{2}$$

式中,$Z(o)$ 为归一化因子,保证了所有可能的状态序列的概率和为 1,即:

$$Z(o)=\sum_{r}\exp\Big(\sum_{i,k}\lambda_k t_k(r_{i-1},r_i,o,i)+\sum_{i,l}\mu_l s_l(r_i,o,i)\Big). \tag{3}$$

式中,t_k 和 s_l 是特征函数,λ_k 和 μ_l 是对应的权值[13-14].

1.2　Bi-LSTM-CRF 模型

长短期记忆网络(long short-term memory,LSTM)是一种解决序列标注中出现的长依赖问题的 RNN 模型. 它与一般的 RNN 结构上并没有本质的区别,只是使用了不同结构的隐藏单元结构,在 LSTM 中存在一个被称为 Cell 的结构. 它包含 1 个细胞状态和 3 个门,即输入门、忘记门、输出门,并通过这 3 个门来控制细胞状态[15],如图 2 所示. 其中,输入门决定保留当前输入的多少信息,忘记门决定保留上一个隐层传来的多少信息,输出门决定将输出多少的信息. 图 2 中 x_t 为输入序列,i_t 和 o_t 分别为保留下来的输入、输出序列,h_t 为细胞的输出序列,C_t 为细胞的状态,f_t 为激活函数. 每个门通过 sigmoid 层和 pointwise 层的操作来对输入到门的信息进行选择和删除[16].

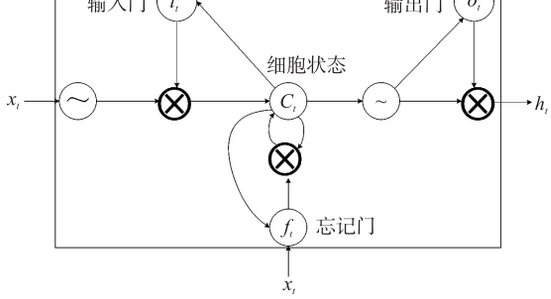

图 2　LSTM 细胞结构图

Fig. 2　LSTM cell structure diagram

Bi-LSTM 则是对 LSTM 进行了一定程度上的优化. Bi-LSTM 通过向前和向后处理每个序列,分别捕获过去和未来信息的两个单独的隐藏状态,将两个隐藏状态连接起来以形成最终输出,以此来更加有效地利用上下文信息,进而输出给定输入句子的最佳标记链.

本文使用句子作为神经网络的输入. 将每句内容转化为词向量,并对于句子中的每个单词进行了单词嵌入. 另外,本文将这些单词在给定句子中的单词序列中嵌入到双向 LSTM 网络中,其中计算每个单词的前向和后向表示. 最后,将两个矢量连接并馈送到 CRF 层,以联合解码最佳标签序列并获得每个字的预测,如图 3 所示[17]. 其中自前向后循环神经网络层的更新公式为:

$$\vec{\boldsymbol{h}}_i = H(W_{x\vec{h}_i}x_t + W_{\vec{h}\vec{h}}\vec{\boldsymbol{h}}_{i-1} + b_{\vec{h}_i}). \tag{4}$$

自后向前循环神经网络层的更新公式为:

$$\overleftarrow{\boldsymbol{h}}_i = H(W_{x\overleftarrow{h}_i}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{\boldsymbol{h}}_{i+1} + b_{\overleftarrow{h}_i}). \tag{5}$$

两层循环神经网络层叠加后输入隐藏层:

$$P_i = W_{\vec{h}y}\vec{\boldsymbol{h}}_i + W_{\overleftarrow{h}y}\overleftarrow{\boldsymbol{h}}_i + b_y. \tag{6}$$

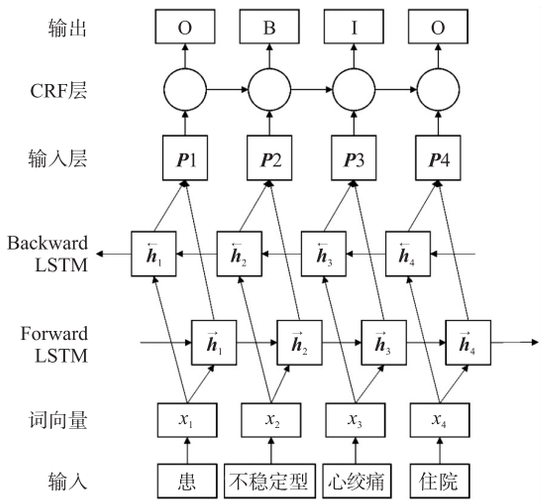

图 3　Bi-LSTM-CRF 模型

Fig. 3　Bi-LSTM-CRF model

2　实验

2.1　数据来源

本文从河南省某医院获取了 108 份冠心病真实门诊病历作为研究数据,病历文本共计 43 736 字符. 其中大部分病历数据内容按照类型可分为 5 个组成部分.

(1)就诊历史:包括患者过去几年中身体出现的症状,采取的措施,曾被诊断为哪些疾病,是否经过治疗,在哪些地方经受过何种治疗,治疗效果如何等信息的记录.

(2)本次就诊原因:记录了患者来院就诊之前的身体状况和出现的症状,以及症状的发作频次和发作程度. 多为患者自述.

(3)具体检查诊断结果:该部分记录了患者就诊后详细的检查情况,并记录了医生对患者的诊断结果. 部分病历还记录了具体的治疗措施和方法.

(4)入院症状:详细记录了患者的发病症状和身体情况,包括饮食状况,睡眠状况,意识状况和二便状况.

(5)病史:该部分记录了患者的先前所患的疾病,包括:疾病名称,患病时间和程度,治疗或处理的手段,目前的状况.

其中小部分病历缺少就诊历史或者病史. 同时对于病历中每个部分的具体内容记录上也存在缺失. 本文使用了序列标注中常用的 BIO 标注法,在这种标注下,实体可能被切割成几个部分,也可能单独出现,如:“冠心病 B”“慢性 B 胃炎 I”或者是“不 B 稳定型 I 心绞痛 I”. 并在实验过程中通过算法来准确界定实体的边界,从而将实体准确地识别出来.

2.2　评价标准

本文的主要任务是在病历文本中尽可能识别和抽取出医学命名实体. 因此本文使用精确率、召回率和 $F1$ 值作为评测指标,同时在测评过程中使用了十字交叉验证法(10-fold cross-validation). 该种测试方法优点在于可以评估模型是否存在过拟合问题,同时评估模型的泛化能力.

2.3　实验结果

本文针对 CRF 模型选取了不同的特征组成不同的特征集,进行训练和测试,并把测试结果与标准数据答案集进行匹配比较,分析评价性能. 本文实验中,将词特征、转移特征作为基本特征,用 Baseline 来表示. 词性特征、上下文关键特征,构词特征作为备选组合特征,并通过选取不同特征模板组合来分析不同特征对实验结果的影响. 本文选取的特征组合如表 1 所示,$X1$、$X2$、$X3$、$X4$ 分别表示词性特征、上文关键词特征(窗口为[-3,0]),构词特征,下文关键词特征(窗口为[0,3]). 其中上下文关键词特征的窗口距离为 3.

表 1　特征集

Table 1　Feature template set

特征	特征选取
Baseline	Baseline
$M1$	Baseline+$X1$
$M2$	Baseline+$X1$+$X2$
$M3$	Baseline+$X1$+$X2$+$X3$
$M4$	Baseline+$X1$+$X2$+$X4$
$M5$	Baseline+$X2$+$X3$+$X4$
$M6$	Baseline+$X1$+$X2$+$X3$+$X4$

通过表 2 可以看出,Baseline 特征集在精确率指标最高,召回率和 $F1$ 值低. 随着特征的不断加入,精确率变化整体呈下降趋势,召回率和 $F1$ 值皆有所提高. 其中结合词性特征、上文关键词特征、构词特征、下文关键词特征的 $M6$ 特征集召回率值和 $F1$ 值最高. 因此可以得出结合词性特征、上文关键词特征、构词特征、下文关键词特征的特征集在不同特征集的比较上抽取性能最高. 不同特征集实体识别性能如图 4 所示.

Bi-LSTM-CRF 模型不依赖于手工设计特征集,因此不存在不同特征集之间的比较,表 3 展示了 CRF 模型和 Bi-LSTM-CRF 模型实验结果的比较. 从表 3 中可以明显看出手工设计特征集的 CRF 模型无论是在精确率、召回率还是 $F1$ 值上都要优于 Bi-LSTM-CRF 模型. 其中,CRF 模型比较于 Bi-LSTM-CRF 模型在精确率上高 13.57%,召回率上高 8.52%,$F1$ 值上高 10.63%,指标差距较为明显. 不同模型之间的识别性能如图 5 所示.

根据上述实验,本文基于 CRF 模型研发了实体识别系统,如图 6 所示.

表 2　测试结果

Table 2　Test results

特征集	精确率	召回率	$F1$ 值
Baseline	95.24%	74.28%	83.04%
$M1$	93.85%	74.55%	82.70%
$M2$	92.86%	78.11%	84.60%
$M3$	94.22%	81.93%	87.40%
$M4$	92.38%	78.73%	84.74%
$M5$	94.13%	81.43%	87.09%
$M6$	93.81%	82.61%	87.63%

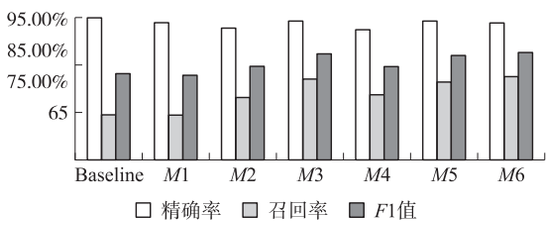

图 4　不同特征模板性能比较

Fig. 4　Performance comparison of different feature templates

表 3　不同模型性能比较

Table 3　Comparison of performance of different models

模型	精确率	召回率	$F1$ 值
CRF	93.81%	82.61%	87.63%
Bi-LSTM-CRF	80.24%	74.09%	77.00%

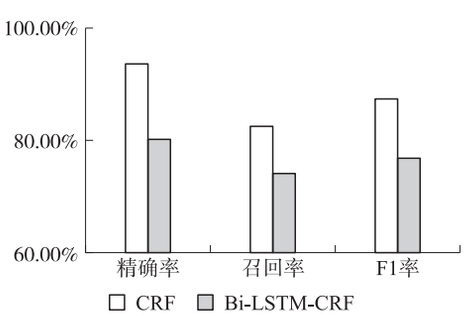

图 5　不同模型性能比较

Fig. 5　Comparison of different model performance

图 6　实体识别系统

Fig. 6　Entity identification system

3　结论

本文所使用的数据为河南省某医院冠心病真实门诊病历. 门诊病历相较于住院病历在内容上更加精炼,每份病历的内容较短,结构组成单一,每份病历的结构内容大体相似. 疾病实体具有个数少、重复率高的特点,因此疾病实体的特征十分明显. 通过选择不同特征集之间的实验结果的对比,认为词性组合在实体构成上存在规律,并且通过加入语义特征和构词特征,可以在略微损失精确率的情况下得到更高的召回率和 $F1$ 值. 同时本文进一步比较了 CRF 模型和 Bi-LSTM-CRF 模型的性能. 其中,栗伟[7]使用基于规则的 CRF,针对测试集获得了 $F1$ 值为 89%,杨红梅[10]使用 Bi-LSTM 针对测试集获得了 $F1$ 值 80%左右. 本文所使用的测试数据集中,CRF 针对测试集的 $F1$ 值为 87.63%,Bi-LSTM-CRF 的 $F1$ 值为 77.00%,CRF 模型性能要高于 Bi-LSTM-CRF 模型. 根据实验结果显示,本文认为在真实领域门诊病历中,由于特征十分显著,手工设计特征集的 CRF 模型相比于运用深度学习的 Bi-LSTM-CRF 模型具有更好的性能.

从真实病历出发,选取了 108 份冠心病真实门诊病历文本作为数据. 分别使用 CRF 和 Bi-LSTM-CRF

模型,对这些病历进行了疾病命名实体的抽取. 经过训练和测试评估之后,发现 CRF 模型更适用于真实门诊病历文本实体抽取. 通过使用 CRF 模型对该类病历文本中的疾病实体实现了较好的抽取,为后续的实体间关系的抽取和病历文本的知识发现打下了坚实的基础.

[参考文献](References)

[1] DICK R S, STEEN E B, DETMER D E. The computer-based patient record: an essential technology for health care[M]. Washington DC: National Academy Press, 1997.

[2] 李彬. 电子病历的应用现状及发展对策初探[J]. 医学与社会, 2005, 18(6): 46-49.

[3] 李丽双, 何红磊, 刘珊珊, 等. 基于词表示方法的生物医学命名实体识别[J]. 小型微型计算机系统, 2016, 37(2): 302-307.

[4] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.

[5] TONG F, LUO Z H, ZHAO D S. A deep network based integrated model for disease named entity recognition[C]//Proceedings of IEEE International Conference on Bioinformatics and Biomedicine. Washington DC: IEEE Computer Society, 2017: 618-621.

[6] 李莹. 文本病历信息抽取方法研究[D]. 杭州: 浙江大学, 2009.

[7] 栗伟. 电子病历文本挖掘关键模型研究[D]. 沈阳: 东北大学, 2014.

[8] 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.

[9] 张晓斌. 基于 CNN 和双向 LSTM 融合的实体关系抽取[J]. 网络与信息安全学报, 2018, 4(9): 44-51.

[10] 杨红梅. 基于双向 LSTM 神经网络电子病历命名实体的识别模型[J]. 中国组织工程研究研究, 2018, 32(4): 1082-1086.

[11] 曹春萍. 基于 E-CNN 和 BLSTM-CRF 的临床文本命名实体识别[J]. 计算机应用研究, 2019, 36(12): 3748-3751.

[12] 栗伟. 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, 22(20): 3237-3242.

[13] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[14] 刘凯. 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.

[15] 柏兵. 基于 CRF 和 BI-LSTM 的命名实体识别方法[J]. 中国信息科技大学学报, 2018, 33(6): 27-33.

[16] 陈彦妤. 基于 CRF 和 Bi-LSTM 的保险名称实体识别[J]. 智能计算机与应用, 2018, 8(3): 112-114.

[17] 金宸. 基于双向 LSTM 神经网络模型的中文分词[J]. 中文信息学报, 2018, 32(2): 29-37.

[责任编辑:陈　庆]