本文引用格式：陈逸墨,叶辉,易珺,等.基于 BERT-BiLSTM-CRF 模型的电子病历隐私信息识别方法[J].自动化与信息工程,2022,43(2):35-40.

CHEN Yimo, YE Hui, YI jun, et al. Private information recognition method of electronic medical records based on Bert-BiLSTM-CRF model[J]. Automation & Information Engineering, 2022,43(2):35-40.

基于 Bert-BiLSTM-CRF 模型的电子病历隐私信息识别方法*

陈逸墨[1]　叶辉[1]　易珺[2]　周华文[1]　方丹丹[1]　曹东[1]

（1.广州中医药大学医学信息工程学院，广东 广州 510006

2.广东药科大学医药信息工程学院，广东 广州 510006）

摘要：随着电子病历数据开放共享的需求越来越大，电子病历去隐私性问题亟需解决。利用自然语言处理技术，提出一种基于 Bert-BiLSTM-CRF 模型的电子病历隐私信息识别方法。采用某三甲中医院的电子病历作为数据来源，结合当前公开的数据集进行训练，得到正确率为 94.02%、召回率为 94.25%、*F*1 为 93.98%的中医电子病历隐私信息识别模型。与其他传统模型进行对比实验表明，Bert-BiLSTM-CRF 模型能有效识别并保护电子病历中的隐私数据，有助于医疗数据的开放共享。

关键词：隐私信息；Bert；双向长短时记忆网络；条件随机场；电子病历

中图分类号：TP391.1　　**文献标识码：**A　　**文章编号：**1674-2605(2022)02-0006-06

DOI：10.3969/j.issn.1674-2605.2022.02.006

0　引言

随着信息时代的到来，各行各业开始与互联网、信息技术交融并飞速发展。为加快我国医学领域的发展进程，卫生部发布了《电子病历基本规范（试行）》。电子病历也称计算机化的病案系统，是用电子设备保存、管理、传输和重现的数字化医疗记录[1]。电子病历在提高信息交流效率的同时也面临诸多挑战，其中如何有效识别并隐藏患者的隐私信息成为关键问题。

目前，中文命名实体识别方法主要基于规则、统计机器学习和深度学习等方法[2]。其中，基于规则的方法依赖手工规则，结合命名实体库，通过实体与规则的相符情况进行类型判断。该方法能够得到较好的识别效果，但不同领域的规则各有不同且这些规则不能互用，因此机器学习的方法逐渐兴起。目前，用于中文命名实体识别的机器学习模型主要有隐马尔科夫模型（Hidden Markov model, HMM）、条件随机场（condition random field, CRF）[3]等。随着硬件计算能力的提升，基于深度学习的方法越来越普遍，且效果较基于统计机器学习的方法更胜一筹。目前，基于深度学习的方法主要通过神经网络来训练模型，主流神经网络模型有卷积神经网络（convolutional neural networks, CNN）[4]、循环神经网络（recurrent neural network, RNN）[5]、长短时记忆神经网络（long short–term memory, LSTM）[6]等。中医电子病历具有复杂性高、词语多义性强、专业性强等特点，传统模型虽然可以实现实体识别功能，但效果不尽如人意。

近年来提出的 Bert 预训练语言模型，凭借优秀的表意能力，使与之结合的神经网络模型效果更佳。本文提出由 Bert、BiLSTM 和 CRF 三个模块构成的模型对中医电子病历中的隐私信息进行识别。

1　模型原理

Bert-BiLSTM-CRF 模型框架如图 1 所示。

首先，待处理的数据输入 Bert 进行预训练；然后，BiLSTM 层进行语义编码处理；最后，将得到的数据输入 CRF 层计算最终结果。与传统的基于深度学习方法相比，本文方法引入了 Bert 预训练语言模型。

* 基金项目：国家重点研发计划资助（2019YFC1710400）；广东省普通高校重点领域专项（2020ZDZX3080）。

Bert-BiLSTM-CRF 模型是经过大量语料及长时间训练得到的，能根据上下文信息计算出字的向量表示，可有效表现字的多义性，增强句子的语义表示[7]。

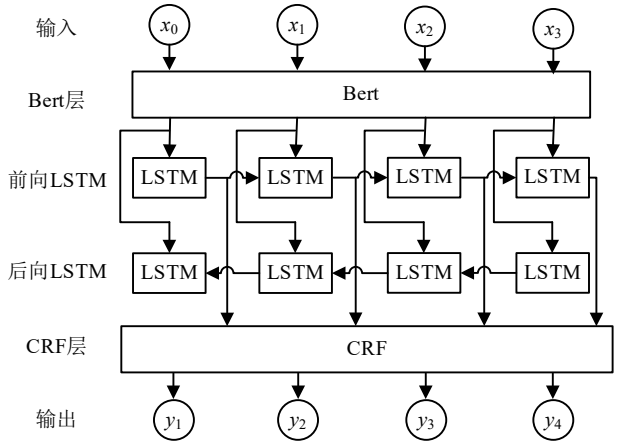

图 1　Bert-BiLSTM-CRF 模型示意图

1.1　Bert 预训练语言模型

2015 年，DAI 和 LE 首次提出预训练语言模型[8]。2018 年 DEVLIN 等经过改进，提出 Bert 预训练语言模型[9]。该模型的构成元素为表义能力较强的 Transformer[10]。Transformer 是一种基于 Attention 机制的深度网络，具有良好的并行计算能力且善于捕捉长距离特征，结构如图 2 所示。

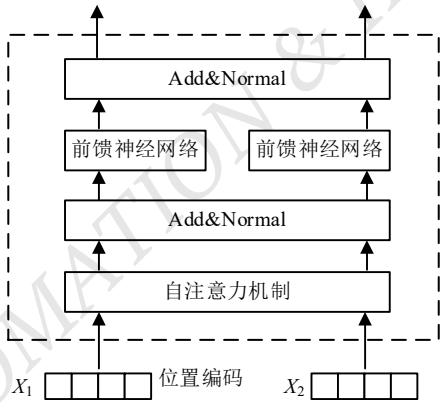

图 2　Transformer 编码单元

首先，将预处理的数据输入 Bert 预训练语言模型；然后，Bert 利用具有特殊作用的标志划分句子，如“患者生于广东省广州市，久居本地”会被 Bert 转为“[CLS]患者生于广东省广州市，[SEP]久居本地”，其中[CLS]标志放在第一个句子首位，[SEP]标志用于分开 2 个输入句子；最后，Bert 将句子转换为 Embedding，输出词向量$\vec{\boldsymbol{B}}$。

1.2　BiLSTM

1997 年，HOCHREITER 提出基于 RNN 改进的 LSTM[11]。LSTM 模型较于 RNN 模型具有可利用长距离信息的特点，并解决了 RNN 模型存在的梯度消失问题。2005 年，GRAVES 根据 LSTM 和双向 RNN 提出双向长短时记忆网络（BiLSTM）[12]。LSTM 单元主要由输入门、遗忘门、输出门 3 部分组成。其中，输入门确定保留信息；遗忘门确定丢弃信息；输出门确定可输出信息，结构图如图 3 所示。

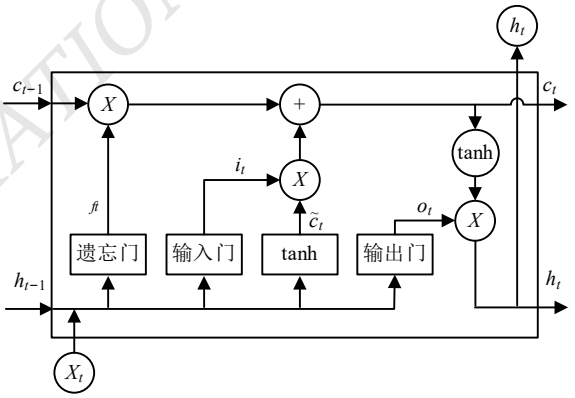

图 3　LSTM 单元结构图

首先，将 Bert 预训练语言模型中输出的某一词向量$\vec{\boldsymbol{B}}$输入到 BiLSTM，X_t 表示 t 时刻的输入数据，$\overrightarrow{h}_t$ 和 $\overleftarrow{h}_t$ 分别表示 t 时刻正向 LSTM 与反向 LSTM 的输出$\overrightarrow{h}_t=(\overrightarrow{h}_1,\overrightarrow{h}_2,\dots,\overrightarrow{h}_t)$、$\overleftarrow{h}_t=(\overleftarrow{h}_1,\overleftarrow{h}_2,\dots,\overleftarrow{h}_t)$；然后，将两者组合得到 BiLSTM 在 t 时刻的输出 $h_t=[\overrightarrow{h}_t,\overleftarrow{h}_t]$，从而得到一个隐状态序列$(t_1,t_2,\dots,t_i)$；最后，给每一个数据带上一个预测分值输入 CRF 层。

1.3　CRF

CRF 是 LAFFERTY 等在 2001 年提出的一种判别式模型，属于随机场的一种[13]。由于 BiLSTM 模块的输出结果会出现一些无意义的字符和标签，也没有

考虑其间的依赖关系，故通过加入 CRF 模块从训练数据中获得约束性规则，保证标签是合法的[14]。BiLSTM 模块的输出序列作为 CRF 模块的输入，如句子 $X(x_1, x_2, ..., x_i)$通过 Bert 预训练语言模型和 BiLSTM 层后，得到每个字的预测序列 $Y(y_1, y_2, \ldots, y_i)$，进入 CRF 后被转换为 BIO 标注法所定义的标记 $Tag(tag_1, tag_2, \ldots, tag_j)$，其中 j 表示标记维度。通过 CRF 层为标记打分，采用 Softmax 函数进行归一化，以 BIO 标注法对标记序列进行规整，完成隐私信息的识别[15]。

2 实验

本实验用来识别中医电子病历中的隐私信息，包括人名、地名、机构名、年龄。实验数据主要来自人民日报语料库和某三甲中医院的电子病历，其中电子病历 349 份，共 11 465 469 个字。将人民日报语料库与电子病历中的数据以 1∶9 的比例分割后作为测试集和训练集。为保证数据整洁，人民日报语料库已标记的数据不做变动，在电子病历数据中新增“年龄”实体类型，用以识别年龄信息。实验主要分为数据预处理、数据导入模型、评判结果 3 个步骤。

2.1 数据预处理

本文所用数据均采用 BIO 标注法，标签有 9 种：B-PER、I-PER、B-LOC、I-LOC、B-ORG、I-ORG、B-AGE、I-AGE、O。其中，B 表示实体开始部分；I 表示实体非开始部分；O 表示非实体；PER 表示人名实体；LOC 表示地名实体；ORG 表示机构实体；AGE 表示年龄实体。BIO 标签集如表 1 所示。

表 1 BIO 标签集

实体类型	开始标记	中间和结尾标记
人名	B-PER	I-PER
地名	B-LOC	I-LOC
机构名	B-ORG	I-ORG
年龄	B-AGE	I-AGE
非实体标记	O	O

未标记的数据使用自主开发的标注软件进行实体标注，操作界面如图 4 所示，标注结果如图 5 所示。其中，C 为需要标注的实体；P 为实体在文本中的位置；T 为实体类型。

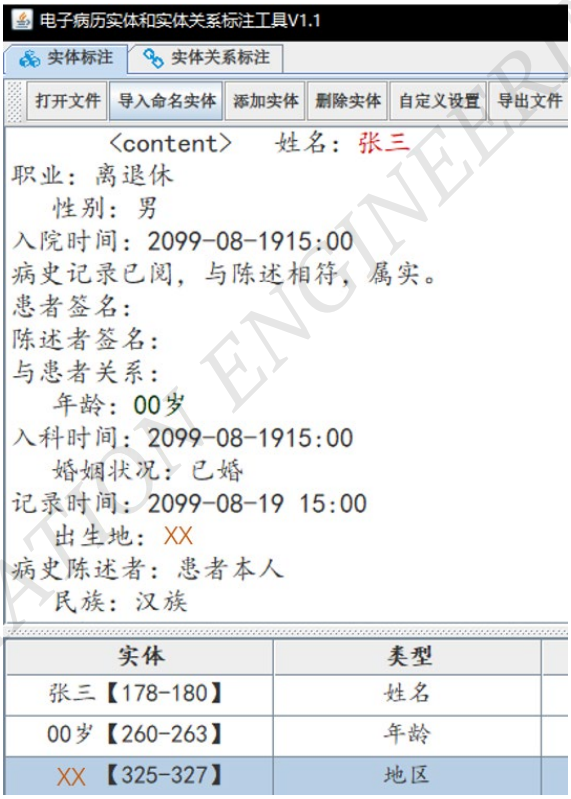

图 4 标注软件操作界面

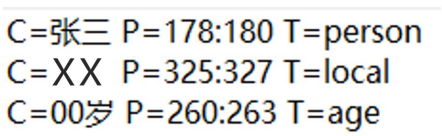

图 5 标注结果

2.2 实验参数设置

3 个模型的常规参数设置如表 2 所示。其中“Max sequence length”表示字数个数；“epoch”表示时期，一个 epoch 就是将所有训练样本训练一次的过程；“dropout”表示防止过拟合参数；“Learning rate”表示学习率，合适的学习率可以使目标函数在合适的时间内收敛到局部最小值；“Batch size”表示每批样本的大小；“Max checkpoints”表示训练过程中的最大

模型快照。

表 2　实验参数

参数	取值
Max sequence length	300
epoch	30
dropout	0.5
Learning rate	0.001
Batch size	20
Max checkpoints	3

2.3　实验结果

本文以正确率 P、召回率 R、和 $F1$ 值作为评判指标。其中，正确率 P 为预测准确样本中真实准确样本的比例；召回率 R 为真实准确样本中预测准确样本的比例；$F1$ 为正确率和召回率的加权平均值。

$$P=\frac{\text{识别出的正确实体个数}}{\text{识别出的所有实体个数}}\times 100\%$$

$$R=\frac{\text{识别出的正确实体个数}}{\text{所有标注的实体个数}}\times 100\%$$

$$F1=\frac{2PR}{P+R}\times 100\%$$

实验环境如表 3 所示。

表 3　实验环境

类别		配置
硬件	GPU	GTX 2080
	CPU	E5-2650L V3 8 核
	操作系统	Ubuntu 18.04 Linux 64 位
	内存	64 GB
	显存	11 GB GDDR6
软件	Python	3.6.12
	Tensorflow	2.2.0
	CUDA	11.0

本文实验中所有模型均在上述配置下完成训练。各实体类型的识别结果如表 4 所示。

表 4　3 种模型对不同实体类型的识别结果

模型	实体类型	P	R	$F1$
BiLSTM	PER	0.823	0.831	0.827
	ORG	0.781	0.788	0.784
	LOC	0.831	0.835	0.833
	AGE	0.792	0.80	0.796
BiLSTM-CRF	PER	0.875	0.854	0.856
	ORG	0.802	0.814	0.796
	LOC	0.864	0.871	0.863
	AGE	0.821	0.819	0.82
Bert-BiLSTM-CRF	PER	0.966	0.97	0.967
	ORG	0.907	0.921	0.91
	LOC	0.957	0.956	0.955
	AGE	0.931	0.923	0.927

其中，模型耗时 BiLSTM 为 158.771 min；BiLSTM-CRF 为 336.951 min；Bert-BiLSTM-CRF 为 1 718.366 min。

2.4　实验结果分析

训练后得到平均正确率为 94.02%、平均召回率为 94.25%、平均 $F1$ 为 93.98%的中医电子病历隐私信息识别模型。

从模型方面来看：Bert-BiLSTM-CRF 模型的平均正确率达到 94.02%，在 4 个实体类型上的识别效果都优于 BiLSTM 模型和 BiLSTM-CRF 模型；由此可见，Bert-BiLSTM-CRF 模型比传统的 LSTM 模型效果更好。

从实体类型来看：PER 和 LOC 的识别效果较好，特别是 Bert-BiLSTM-CRF 模型对这 2 种实体类型识别的 $F1$ 值均超过了 0.95，这是由于人民日报语料中人名和地名的标注质量较高且这些实体不会因其他客观因素而改变；ORG 和 AGE 的识别效果较差，主要原因是 ORG 有时用缩略词或组合词对识别产生干

扰，如“广州中医药大学”被缩略为“广中医”、“中山大学第三附属医院”被缩略为“中大三附院”；年龄实体由数字组成，而病例中存在其他与年龄无关的数字，导致 AGE 识别不准确，如“药品剂量 50 mg/1 日”中“50”被识别成年龄，“日期 2011-11-23”中“11”和“23”被识别为年龄，产生信息混淆。

3 结语

本文提出的 Bert-BiLSTM-CRF 模型已达到可以使用的水平，相比传统的 BiLSTM 模型和 BiLSTM-CRF 模型，本文模型识别不同类别隐私信息的能力更强。陈衍旭[16]提出的 Bert-BiLSTM-CRF 模型的隐私信息识别 $F1$ 值为 0.932 9，本文模型在此基础上有一定程度的提升。在之后的工作中，需要丰富数据集并且对模型进行适当改进，以提高模型的识别效率。如明确年龄实体与其他包含数字的实体的分类；通过增加原始数据数量来增加训练量。近年来有融入注意力机制[17]的新模型出现。因此，下一步可以考虑从数据处理和融入新机制入手来提升模型性能。

参考文献

[1] 中华人民共和国卫生部.电子病历基本规范（试行）[J].中国药房,2010,21(12):1063-1064.

[2] 刘浏,王东波.命名实体识别研究综述[J].情报学报,2018,37(3):329-340.

[3] 宫义山,段亚奇.基于不同模型的中文命名实体识别方法研究[J].长江信息通信,2021,34(1):84-86.

[4] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989,1(4):541-551.

[5] 夏瑜潞.循环神经网络的发展综述[J].电脑知识与技术,2019,15(21):182-184.

[6] HOCHREITER S, SCHMIDHUBER J. LSTM can solve hard long time lag problems[J]. Advances in neural information processing systems, 1997: 473-479.

[7] 王远志,曹子莹.Bert-BLSTM-CRF 模型的中文命名实体识别[J].安庆师范大学学报(自然科学版),2021,27(1):59-65.

[8] DAI A M, LE Q V. Semi-supervised sequence learning[J]. Advances in neural information processing systems, 2015,28: 3079-3087.

[9] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[10] VASWANI A, SHAZEER N, PARMAR N. et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, 2017:5998-6008.

[11] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[12] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.

[13] LAFFERTY J, MCCALLUM A, PEREIRA F. Condition random fields: probabilistic models for segmenting and labeling sequence data[C].Proceedings of the 18th International Conference on Machine Learning, 2001, 951:282-289.

[14] 罗熹,夏先运,安莹,等.结合多头自注意力机制与 BiLSTM-CRF 的中文临床实体识别[J].湖南大学学报(自然科学版),2021,48(4):45-55.

[15] 刘一斌,叶辉,易珺,等.基于朴素贝叶斯和 word2vec 的中医电子病历文本信息抽取[J].世界科学技术-中医药现代化,2020,22(10):3563-3568.

[16] 陈衍旭.面向临床文本的知识获取与应用[D].哈尔滨工业大学,2019.

[17] 张华丽,康晓东,李博,等.结合注意力机制的 Bi-LSTM-CRF 中文电子病历命名实体识别[J].计算机应用,2020,40(S1):98-102.

Private Information Recognition Method of Electronic Medical Records Based on Bert-BiLSTM-CRF Model

CHEN Yimo[1] YE Hui[1] YI jun[2] ZHOU Huawen[1] FANG Dandan[1] CAO Dong[1]

(1.School of Medical Information Engineering, Guangzhou University of Chinese Medicine,

Guangzhou 510006, China 2. College of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, China)

Abstract: With the increasing demand for open sharing of electronic medical record data, the privacy of electronic medical record needs to be solved urgently. Using natural language processing technology, a privacy information recognition method of electronic medical record based on Bert-BiLSTM-CRF model is proposed. Using the electronic medical record of a three-tier traditional Chinese medicine hospital as the data source, combined with the current public data set for training, we get the privacy information recognition model of traditional Chinese medicine electronic medical record with the accuracy rate of 94.02%, the recall rate of 94.25% and *F*1 of 93.98%. Compared with other traditional models, the experiment shows that Bert-BiLSTM-CRF model can effectively identify and protect the private data in EMR, which is conducive to the open sharing of medical data.

Keywords: privacy information; Bert; BiLSTM; CRF; electronic medical record

作者简介：

陈逸墨，男，1997 年生，在读研究生，主要研究方向：医学自然语言处理。

叶辉，男，1978 年生，硕士，讲师，主要研究方向：医学自然语言处理。

易珺，女，1976 年生，硕士，副教授，主要研究方向：医学自然语言处理。

周华文，男，1997 年生，在读研究生，主要研究方向：医学自然语言处理。

方丹丹，女，1998 年生，在读研究生，主要研究方向：医学自然语言处理。

曹东（通信作者）男，1975 年生，博士研究生，教授，主要研究方向：医学自然语言处理、医学信号传感与检测。E-mail: caodong@ gzucm.edu.cn

（上接第 34 页）

[33] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]//Proceedings of the IEEE international conference on computer vision, 2017:2980-2988.

Sleep Apnea Detection Method Based on Dilated Convolution and Attention Mechanism

ZHENG Heyu LIN Meina

(Guangdong University of Technology, Guangzhou 510006, China)

Abstract: In order to improve the accuracy of sleep apnea detection based on ECG signal, aiming at the problem that the existing detection methods generally need more complex feature engineering and manual correction steps, can not adaptively preprocess ECG signal and will lose more information, a sleep apnea detection method based on dilated convolution and attention mechanism is proposed. Firstly, the adaptive preprocessing network is used to filter the redundant information in ECG signal (including baseline drift, EMG interference, etc.); Then, the detection network based on dilated convolution and time attention mechanism is used to extract timing features from ECG signal and detect them. The experimental results on Apnea-ECG data set show that compared with the existing detection methods, this method can achieve more effective sleep apnea detection.

Keywords: ECG signal; sleep apnea syndrome; adaptive signal preprocessing; dilated convolution; attention mechanism

作者简介：

郑和裕，男，1996 年生，硕士研究生，主要研究方向：模式识别，机器学习，生物信号处理。E-mail: zheng_hy1209@qq.com

林美娜，女，1997 年生，硕士研究生，主要研究方向：模式识别，生物信号处理。E-mail: meina.lin@mail.gdut.edu.cn