中文电子病历命名实体识别方法研究*

马欢欢 孔繁之 高建强

（曲阜师范大学软件学院 曲阜 273100） （济宁医学院医学信息工程学院 日照 276826）

〔摘要〕 针对中文电子病历命名实体识别任务中存在的边界划分不准确、实体识别率不高等问题，提出基于深度学习的 CNN – BiLSTM – CRF 模型，详细阐述模型结构与原理，采集 3 127 份中文电子病历数据进行实验以验证模型性能，结果表明该模型具有较好的识别效果及性能。

〔关键词〕 中文电子病历；命名实体识别；卷积神经网络

〔中图分类号〕 R – 056 〔文献标识码〕 A 〔DOI〕 10. 3969/j. issn. 1673 – 6036. 2020. 04. 005

Study on Named Entity Recognition Method of Chinese Electronic Medical Records *MA Huanhuan, School of Software, Qufu Normal University, Qufu 273100, China; KONG Fanzhi, GAO Jianqiang, School of Medical Information Engineering, Jining Medical University, Rizhao 276826, China*

〔**Abstract**〕 Aiming at the problems of inaccurate boundary division and low entity recognition rate in the Named Entity Recognition (NER) task of Chinese Electronic Medical Records (EMR), the paper proposes a CNN – BiLSTM – CRF model based on deep learning, expounds the structure and principle of the model in detail, and collects 3 127 Chinese EMR for experiments to verify the performance of the model. The results show that this model achieves better recognition effect and better performance.

〔**Keywords**〕 Chinese Electronic Medical Records (EMR); Named Entity Recognition (NER); Convolutional Neural Network (CNN)

1 引言

1.1 相关概念

命名实体识别（Named Entity Recognition, NER）是自然语言处理领域一个重要的研究方向，在 1995 年正式提出[1]，是信息抽取、信息检索、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分[2]，其任务是识别出待处理文本中 3 大类（实体类、时间类和数字类）、7 小类（人名、机构名、地名、时间、日期、货币和百分比）命名实体。电子病历（Electronic Medical Records, EMR）是医务人员在医疗活动过程中使用医疗机构信息系统生成的文字、符号、图表、图形、数据、影像等数字化信息，能实现存储、管理、传输和重现的医疗记录，是病历的一种记录形式[3]。中文电子病历命名实体识别可自动化地提取出疾病、症状、治疗等具有语义特征的实体，有助于医学研究者构建临床决策系统，为患者提供更高效、便捷的个性化医疗服务。

〔收稿日期〕 2019 – 09 – 24

〔作者简介〕 马欢欢，硕士研究生；通讯作者：孔繁之，教授。

〔基金项目〕 教育部产学合作协同育人项目“高精度人脸识别技术与教学平台建设研究”（项目编号：201801245011）。

1.2 命名实体识别方法

传统命名实体识别方法有基于规则和基于机器学习两类。基于规则的识别方法[4]依靠领域专家构造领域词典，对于没有出现在词典中的实体则设计一定规则进行识别。基于机器学习的识别算法主要包括：隐马尔可夫模型（Hidden Markov Model，HMM）[5]、条件随机场（Conditional Random Field，CRF）模型等，其中 Lafferty 等[6]提出的 CRF 解决了标注偏置问题，在一段时间内成为处理命名实体识别最常用的方法。自 2006 年 Hinton 等提出深度学习后，其在语音、图像领域的成功应用使得越来越多的深度学习技术被迁移到自然语言处理中，相关深度学习模型表现出很好的性能。一些学者将识别问题看作多分类问题，Roberts[7]等在医疗文本的研究中采用支持向量机（Support Vector Machine，SVM）模型达到分类效果。一些学者将识别问题视为序列标注问题，Finkel[8]等采用 CRF 建立自动标注模型，考虑的特征主要包括词特征、前后缀、词性序列和词形态。Lample[9]等提出长短时记忆网络（Long Short Term Memory，LSTM）+ CRF 模型，具有超过 CRF 的识别性能，最大优点在于无需特征工程，使用向量就能达到很好的效果。Huang[10]等提出双向长短时记忆网络（Bi - directional Long Short Term Memory，BiLSTM）- CRF 模型的命名实体识别，以解决 LSTM 模型带来的前后序列特征相关性问题。

1.3 卷积神经网络优势

卷积神经网络（Convolutional Neural Network，CNN）最大优点是能够简化特征提取过程，直接通过模型训练，从图像中提取出较好的特征[11]。可以实现“端到端”学习，通过局部连接、权重共享和降采样 3 大特点完成对输入数据的特征学习和特征表示[12]。CNN 可以有效降低网络复杂度，减少训练参数数目，使模型对平移、扭曲、缩放一定程度上不变形，具有强鲁棒性和容错能力，且易于训练和优化[13]。如今在自然语言处理任务中也常见到卷积神经网络应用[14]。Chiu[15]等采用卷积神经网络抽取字符级特征来进行通用命名实体识别，在公开的英文命名实体识别数据集上取得较好效果。本文在现有研究基础上将命名实体识别作为序列标注任务，针对命名实体识别方法存在的边界划分不准确、实体识别率不高等问题，从中文电子病历行文方式出发，将 BiLSTM - CRF 作为基准模型，采用字向量进行分布式表示，引入 CNN 提取空间语义信息，以达到精确识别效果。

2 模型结构

2.1 CNN - BiLSTM - CRF 模型

本研究模型框架主要由 3 个部分组成，即 CNN、BiLSTM、CRF 模块。首先对病历文本进行分句和分字处理，由预先训练的字向量将输入的句子转换为字向量序列，然后通过 CNN 模块对每个字的字向量进行卷积和池化操作，以提取文本序列的空间特征信息，之后输入 BiLSTM 模块学习上下文特征信息，最后通过 CRF 模块将 BiLSTM 的输出解码为一个最优预测标记序列。模型框架，见图 1。

2.2 词向量

Word2vec 使用的算法是 Bengio 等在 2001 年提出的神经网络语言模型（Neural Network Language Model，NNLM），后来 Milolvo[16]团队对这一算法做了改进。利用 Word2vec 模型训练出来的词向量能很好地包含词语特征，甚至能通过训练好的词向量来计算词语之间的相似度。Word2vec 网络模型分为两种，一种是 CBOW，另一种是 Skip - gram，模型框架，见图 2、图 3。其中 CBOW 模型是从原始语句推测出目标字词，Skip - gram 则相反。本研究采用中文维基百科数据集，使用 Skip - gram 预先训练维度为 100 的字向量。

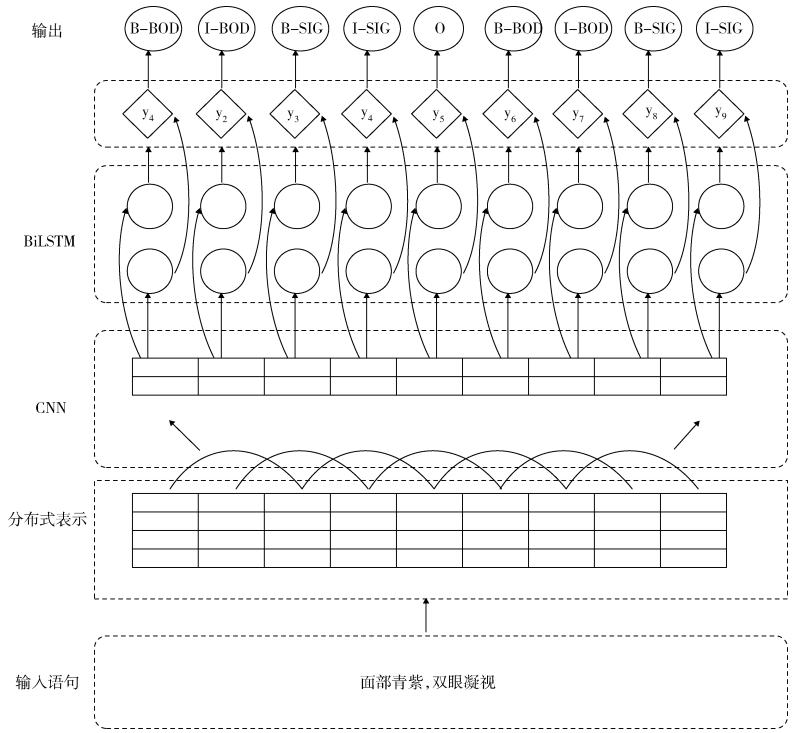

图 1 CNN – BiLSTM – CRF 模型

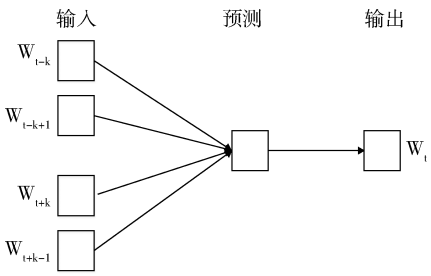

图 2 CBOW 模型

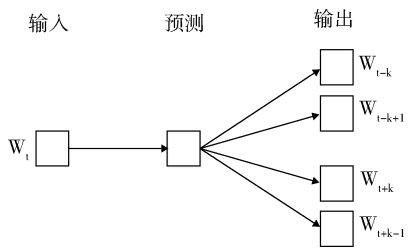

图 3 Skip – gram 模型

2.3 卷积神经网络

CNN 模块主要由输入层、卷积层、池化层和输出层组成，其结构，见图 4。通过分布式表示之后的字向量填充占位符来解决句子长短不一的问题。等长的字向量组成字向量矩阵，首先使用卷积操作处理字向量矩阵，生成多通道特征图，对特征图采用最大池化操作进行下采样得到与卷积核对应的整句话特征，最后输出层输出句子的最终特征表示。卷积层使用 K 个大小为 s 的卷积核在向量序列上以步长 1 进行滑动来提取局部特征。

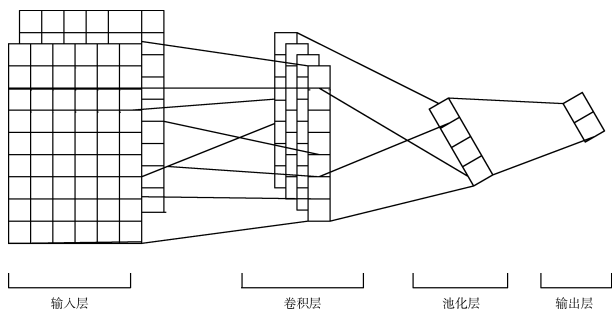

图 4 CNN 模型

2.4 双向长短时记忆网络

循环神经网络（Recurrent Neural Networks，RNN）是一种能够对时序数据进行精准建模的网络。虽然理论上 RNN 能够捕获长距离依赖性，在实践中却由于梯度消失或爆炸而失败。LSTM[17] 是

RNN 的变种，主要应对梯度消失问题。LSTM 模型有 3 个乘法控制单元：输入门（i_t）用来决定在细胞状态（c_t）中存储哪些新信息；遗忘门（f_t）决定将哪些信息从 c_t 中丢弃；输出门（o_t）决定哪些信息作为输出。LSTM 单元在 t 时刻更新的公式如下：

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \quad (1)$$

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \quad (2)$$

$$\tilde{c}_t = \tanh(W c h_{t-1} + U_c x_t + b_c) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (4)$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_0) \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

其中 i_t 为输入门，f_t 为遗忘门，$\tilde{c}_t$ 为新记忆单元，c_t 为最终记忆单元，o_t 为输出门，h_t 为隐藏层，σ 表示神经网络中 sigmoid 激活函数，tanh 表示双曲正切激活函数，$\odot$ 是对应元素点积，x_t 是在 t 时刻的输入词向量，h_t 是在 t 时刻的输出，U_i、U_f、U_c、U_o 是在不同控制门对应输入向量的权重矩阵，W_i、W_f、W_c、W_o 是隐藏层的权重矩阵，而 b_i、b_f、b_c、b_0 是偏差向量。新记忆单元使用当前单词 x_t 和上一时刻隐藏层状态 h_{t-1} 产生当前新信息 c_t。这些门和记忆单元组合起来极大提升 LSTM 处理长序列数据的能力。对于许多序列标记任务，能够访问过去和未来的信息很重要，然而 LSTM 的隐藏状态仅从过去获取信息，对未来一无所知[18]。BiLSTM 将每个顺序和逆序序列向前和向后呈现给两个单独的隐藏状态，以分别捕获过去和未来的信息。最后连接两个隐藏状态作为最终输出。

2.5 条件随机场

CRF[19]是一种用于标注和切分有序数据的条件概率模型，结合最大熵模型和隐马尔可夫模型特点的全局归一化，在考虑连续标签之间依赖关系的情况下找到最佳输出序列。可以将命名实体识别任务转化成序列标注任务。对于给定的观察序列 $X = (x_1, x_2, \ldots, x_n)$，$x_i$ 表示第 i 个字的输入向量。标记序列 y 的概念可以定义为：

$$\exp\left(\sum_j \lambda_j t_j(yi_{-1}, y_i, X, i) + \sum_k \mu_k s_k(y_i, X, i)\right) \quad (7)$$

其中 $t_j(y_{i-1}, y_i, X, i)$ 为概率转移方程，表示输入序列 X 在其标注为 y_{i-1} 和 y_i 之间的转移概率。$s_k(y_i, X, i)$ 为状态函数，表示对于序列 X 其 i 位置的标记为 y_i 的概率，λ_j，μ_k 分别对应相应函数的权重。

3 实验及结果分析

3.1 实验环境、数据集及评价标准

本研究实验环境为 inter@ Corei5CPU@ 3.30GHz * 4，操作系统选用 Ubuntu16.04，模型框架为 Keras，Keras 是以 TensorFlow 为底层的高级的深度学习链接库，编程语言为 Python。实验数据集是由山东省某三甲医院神经内科提供的 2010 – 2017 年 3 127 份癫痫患者电子病历。采用准确率（Precision，P）、召回率（Recall，R）和 $F1$ 值（$F1$ – score，$F1$）3 个指标作为模型的衡量标准。其中准确率是指正确识别的实体数占总识别实体数的比例，召回率是指正确识别的实体数占总实体数的比例，而 $F1$ 值是准确率和召回率的调和平均值。各项指标具体公式如下：

$$P = \frac{n}{M} \times 100\% \quad (8)$$

$$R = \frac{n}{N} \times 100\% \quad (9)$$

$$F1 = \frac{2PR}{P + R} \times 100\% \quad (10)$$

其中 M 代表识别出的实体个数，N 代表测试集中总实体个数，n 代表正确识别的实体数。

3.2 实验设计与分析

3.2.1 数据预处理 对收集的 3 127 份癫痫患者电子病历，在结合癫痫发病原理和相关文献[20]的基础上制定相应标注规范。首先将数据中实体类型分为 3 类，即疾病、症状、治疗。为清楚表示语料中待识别的命名实体，采用 BIO 标记方式来标记实体，其中 B 表示实体的开始，I 表示实体的中间位置，O 表示不属于预分类的实体。然后分别进行数据去敏清洗及分句处理，利用 jieba 分词加载外部用户词典，在领域专家的指导下对病历文本进行自动化标注，完成语料库构建，分配模型训练集、验证集和测试集，分配比例分别是 6∶2∶2。

3.2.2 实验参数设置 本研究所有词向量的维度设置为 100，在训练过程中优化器采用 Adam，学习率设置为 0.1，为减轻过拟合，在 CNN 模块使用

Relu 作为非线性激活函数，同时在 CNN 的输入、BiLSTM 的输入和输出中使用 Dropout，取值为 0.5。

3.2.3 实验结果分析 为验证本研究模型结构的有效性，设计 3 组对比实验，分别与 LSTM、BiLSTM、BiLSTM－CRF 模型进行对比。实验结果，见表1。可以发现 BiLSTM 模型较 LSTM 模型 *F*1 值提高 5.01%。LSTM 模型的神经元信息只能从前向后传递，当前时刻的输入信息仅能利用之前时刻的信息，且命名实体识别的标签之间具有强烈的依赖关系，而 BiLSTM 则既能利用当前又能利用之后时刻的信息，说明 BiLSTM 模型在中文电子病历文本命名实体识别任务中具有良好的适应性。另外 BiLSTM－CRF 模型较 BiLSTM 模型 *F*1 值提高 5.74%。由于 BiLSTM 模型的输出是对于句子中每个单词对应的标签类别的最高预测分数值，但最高分数值往往存在漂移现象，即 B－Dis 后面最高分数值会出现 I－Tre，非常不合理。而 CRF 能够从训练数据中学习标注序列的一些约束，可以为 BiLSTM 模块输出的最终预测标签添加一些约束来确保合理有效，即 B－Dis 后面会出现 I－Dis。本研究提出在 BiLSTM－CRF 基础上添加 CNN 模块，较 BiLSTM－CRF 准确率提高 1.15%，召回率提高 1.19%，*F*1 值提高 1.47%，说明 CNN 提取的病历文本特征可作为上下文特征的补充，通过卷积和池化操作能够进一步提升模型识别效果。

表1 模型验证对比实验结果（%）

模型	*P*	*R*	*F*1
LSTM	75.53	78.42	76.94
BiLSTM	81.26	82.66	81.95
BiLSTM－CRF	87.64	87.75	87.69
CNN－BiLSTM－CRF	88.79	89.54	89.16

4 结语

本研究充分考虑中文电子病历的语言特性，结合词嵌入技术将文字映射成高维向量，通过 CNN 模块对每个字向量进行卷积和池化操作，以提取文本序列的空间特征信息，之后输入 BiLSTM 模块学习上下文特征信息，最后通过 CRF 模块将 BiLSTM 输出解码出一个最优预测标记序列。本研究提出的 CNN－BiLSTM－CRF 模型与 LSTM、BiLSTM、BiLSTM－CRF 模型进行实验对比，*F*1 值为 89.16%，相较于其他模型达到较好的识别效果，对医学领域的实体识别研究具有一定参考价值。由于本研究模型识别的数据来源于 3 127 份电子病历，医学实体种类较少，未涉及电子病历中出现的疾病诊断分类、检查措施、身体部位等，故下一步工作需要进一步丰富模型的识别种类，充分挖掘中文电子病历中的医疗信息，在特征选择多样性方面进一步研究。另外可以尝试训练字词联合的向量表示，挖掘字词之间更加紧密的联系。

参考文献

1 Grishman R, Sundheim B. Message Understanding Conference 6: a brief history [C]. Stroudsburg: Proceedings of the 16th Conference on Computational Linguistics－Volume 1, 1996: 466－471.

2 LIU Liu, WANG Dongbo. A Review on Named Entity Recognition [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37 (3): 329－340.

3 《中国卫生质量管理》编辑部. 电子病历基本规范（试行）[J]. 中国卫生质量管理, 2010, 17 (4): 13－14.

4 周昆. 基于规则的命名实体识别研究 [D]. 合肥: 合肥工业大学, 2010.

5 Rabiner L R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition [J]. Proceedings of the IEEE, 1989, 77 (2): 257－286.

6 Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: probabilistic models for segmenting and labeling sequence data [J]. Proceedings of Icml, 2001, 3 (2): 282－289.

7 Roberts A, Gaizauskas R, Hepple M. Extracting Clinical Relationships from Patient Narratives [C]. Stroudsburg: Proceedings of the 2008 Workshop on Current Trends in Biomedical Natural Language Processing, 2008: 10－18.

8 Finkel J R, Manning C. Joint Parsing and Named Entity Recognition [C]. Stroudsburg: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009: 326－334.

9 Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition [C]. San Diego: Proceedings of NAACL－HLT, 2016.

10 Zhiheng Huang, Wei Xu, Kai Yu. Bidirectional Lstm – crf Models for Sequence Tagging [EB/OL]. [2015 – 08 – 09]. https://arxiv.org/abs/1508.01991.

11 顾孙炎. 基于深度神经网络的中文命名实体识别研究 [D]. 南京: 南京邮电大学, 2018.

12 刘小安, 彭涛. 基于卷积神经网络的中文景点识别研究 [EB/OL]. [2019 – 03 – 08]. http://kns.cnki.net/kcms/detail/Detail.aspx? dbname = CAPJLAST&filename = JSGG20190307002&v.

13 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述 [J]. 计算机学报, 2017, 40 (6): 1229 – 1251.

14 B Hu, Z Lu, H Li, et al. Convolutional Neural Network Architectures for Matching Natural Language Sentences [EB/OL]. [2015 – 03 – 11]. https://arxiv.org/abs/1503.03244v1.

15 Jason P C Chiu, Eric Nichols. Named Entity Recognition with Bidirectional LSTM – CNNs [EB/OL]. [2015 – 11 – 26]. https://arxiv.org/abs/1511.08308.

16 Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality [J]. Advances in Neural Information Processing Systems, 2013 (26): 3111 – 3119.

17 Hochreiter S, Schmidhuber, Jürgen. Long Short – term Memory [J]. Neural Computation, 1997, 9 (8): 1735 – 1780.

18 Ma X, Hovy E. End – to – end Sequence Labeling via Bi – directional Lstm – cnns – crf [C]. Berlin: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1064 – 1074.

19 陈斌, 周勇, 刘兵. 基于卷积双向长短期记忆网络的事件触发词抽取 [J]. 计算机工程, 2019, 45 (1): 153 – 158.

20 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 – 2727.

2020 年《医学信息学杂志》编辑出版重点选题计划

2019 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 2019 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“互联网 + 医疗健康”支撑体系、服务体系建设；2 医药卫生信息化发展规划与战略；3 信息化助力医疗服务、公共卫生服务、医疗保障体系建设的技术方案与典型案例；4 医疗卫生信息标准化与规范化建设现状和应用实践；5 医疗卫生信息化相关法律法规；6 智慧医院及智慧医疗服务模式建设目标、发展规划、解决方案。

二、医学信息技术

1 医疗人工智能及健康智能设备研究与应用；2 健康医疗大数据的管理及应用创新；3 家庭医生签约智能化平台建设及网上签约服务；4 精准医学与个性化医疗技术研究与应用；5 物联网、远程医疗服务与健康管理；6 医疗云平台功能、技术、系统架构及基础设施构建；7 基于互联网技术的医疗联合体建设与信息互通共享；8 网络安全体系建设与风险评估。

三、医学信息研究

1 医学信息学基础理论及方法研究；2 医学科技创新体系和发展战略；3 公民健康素养培养及健康促进；4 医学智库研究与智库服务；5 医药卫生数据分析、挖掘与知识发现技术。

四、医学信息组织与利用

1“互联网 +”环境下医学图书馆的创新举措；2 人工智能技术及其在医学图书馆中的应用；3 数字资源建设与学科服务模式演化与机制；4 区域医疗卫生信息资源整合。

五、医学信息教育

1“互联网 +”环境下医学信息专科、本科、研究生教育及继续教育面临的挑战、改革与实践创新；2 医学信息素养教育；3 网络化、数字化医疗健康教育培训平台及在线课程；4 基于互联网的健康科普知识精准教育；5 国外医学信息学教育的先进理念综述。

（《医学信息学杂志》编辑部）