

基于 LSTM-CRF 命名实体识别技术的研究与应用

张聪品 方 滔 刘昱良

(河南师范大学 计算机与信息工程学院 河南 新乡 453007)

摘 要: 随着深度神经网络的发展,深度学习不仅占据了模式识别等领域的统治地位,而且已应用到自然语言处理的各个方面,如中文命名实体识别。对电子病历中的命名实体进行识别时,构建了内嵌条件随机场的长短时神经网络模型,使用长短时神经网络隐含层的上下文向量作为输出层标注的特征,使用内嵌的条件随机场模型表示标注之间的约束关系。该模型识别出了电子病历中的身体部位、疾病名称、检查、症状和治疗五类实体,准确率达到 96.29%,精确率达到了 91.61%,召回率 96.22%,F 值 93.85。其中症状这一实体类别,精确率达到 96.08%,召回率 98.98%,F 值 97.51。实验结果表明,内嵌条件随机场的长短时记忆神经网络模型在识别中文命名实体方面是有效的,有助于自动抽取中文电子病历中实体之间的关系、构建医疗知识图谱。

关键词: 长短时记忆神经网络;条件随机场;命名实体;电子病历

中图分类号: TP319

文献标识码: A

文章编号: 1673-629X(2019)02-0106-03

doi: 10.3969/j.issn.1673-629X.2019.02.022

Research and Application of Named Entity Recognition Based on LSTM-CRF

ZHANG Cong-pin, FANG Tao, LIU Yu-liang

(School of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)

Abstract: With the development of deep neural network, deep learning not only occupies the dominant position in pattern recognition and other fields, but also has been applied to various aspects of natural language processing, such as Chinese named entity recognition. When recognizing named entities in electronic medical records, we construct a long and short time neural network model with embedded random field. The context vector of the hidden layer of long and short time neural networks is used as the feature of output layer annotation, and the embedded conditional random field model to represent the constraint relationship between the annotations. The model identifies five types of entities, including body parts, disease name, examination, symptom and treatment in the electronic medical record, with accuracy of 96.29%, precision rate of 91.61%, recall rate of 96.22%, and F value of 93.85. For the entity category of symptom, the precision rate reaches 96.08%, recall rate of 98.98%, F value of 97.51. The experiment shows that the proposed model is effective in identifying Chinese named entities, which is helpful for the automatic extraction of the relationship between entities in Chinese electronic medical records and the construction of medical knowledge maps.

Key words: long and short time memory neural network; conditional random field; named entity; electronic medical record

0 引 言

电子病历是指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,并能实现存储、管理、传输和重现的医疗记录^[1],是由医务人员撰写的面向患者个体描述医疗活动的记录。

随着自然语言处理技术的发展,可以从电子病历的文本^[2]中自动提取大量专业医疗知识,构建医疗知

识图谱。如电子病历中,“患者缘于 1 年前无明显诱因出现颈肩部及腰部疼痛、右上肢麻木,入院后进行颈椎 CT 检查:颈椎间盘突出。入院后给予患者颈椎牵引、颈部手法推拿、颈部中药塌渍、颈部微波照射治疗。于今日出院。”在病历中,“颈椎 CT 检查”证实了“无明显诱因出现颈肩部及腰部疼痛、右上肢麻木”的发生;而“患者颈椎牵引”、“颈部手法推拿”、“颈部中药塌渍”、“颈部微波照射治疗”这些治疗使患者的症状消

收稿日期: 2018-02-08

修回日期: 2018-06-13

网络出版时间: 2018-11-15

基金项目: 河南省科技攻关项目(152102310313);河南师范大学专业学位研究生课程案例库建设项目(5101119500706)

作者简介: 张聪品(1968-),女,教授,硕导,研究方向为人工智能基础技术、编译技术;方滔(1991-),男,研究生,研究方向为机器学习。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20181115.1046.012.html>

失了,为了从电子病历里抽取这些相关的医疗知识(即关系抽取),构建医疗知识图谱,首先需要识别出电子病历文本中与患者健康密切相关的各类命名实体,如“腰部”等身体部位、“疼痛”等症状、“颈椎CT检查”等检查手段、“颈椎间盘突出”等疾病名称、“颈部手法推拿”等实施的治疗。身体部位、症状、检查手段、疾病名称、治疗这些概念在电子病历信息抽取研究中被称作命名实体^[3]。

传统的中文实体识别方法有条件随机场、字典法和混合方法^[4]。随着深度神经网络技术的发展,深度神经网络技术已经广泛应用于自然语言处理中,包括中文命名实体识别^[5]。基于深度神经网络的中文命名实体识别模型中,使用神经网络隐含层的上下文向量作为输出层标注的特征,但是神经网络模型却无法表示标注之间的约束关系^[6]。

通过在长短时记忆神经网络(LSTM)模型中内嵌条件随机场(CRF)模型,利用CRF模型表示标注之间的约束关系。构建了LSTM-CRF模型,自动识别出电子病历中的五类中文命名实体:身体部位、疾病名称、检查手段、症状和治疗,为下一步抽取关系信息^[7],构建医疗知识图谱奠定了基础。

1 LSTM-CRF 模型

传统的神经网络输出只依赖于当前的输入,循环神经网络通过使用带自反馈的神经元,能够处理任意长度的序列,解决了传统神经网络解决不了的变长输入和相互依赖的处理任务^[8]。长短时记忆神经网络模型解决了循环神经网络由于梯度爆炸或消失只能学习到短周期的依赖关系问题^[9]。

LSTM模型通过引入一组记忆单元,使得神经网络具有学习遗忘历史信息,用新信息更新记忆单元的功能。在时刻 t ,记忆单元 c_t 记录了到当前时刻为止的所有历史信息,并受三个“门”控制:输入门 i_t 、遗忘门 f_t 和输出门 o_t 。三个门的计算公式如下所示,三个门元素的值在 $[0, 1]$ 之间。

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

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

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t) \quad (3)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1}) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (5)$$

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

其中, x_t 是当前时刻的输入; σ 是logistic函数; V_i, V_f, V_o 是对角矩阵。遗忘门 f_t 控制每一个内存单元需要遗忘多少信息,输入门 i_t 控制每一个内存单元加入多少新的信息,输出门 o_t 控制每一个内存单元输出

多少信息。

LSTM模型工作时,首先由遗忘门层通过sigmoid来控制确定通过记忆单元的信息。根据上一时刻的输出 h_{t-1} 和当前输入 x_t 来产生一个0到1的 f_t 值,以决定是否让上一时刻学到的信息 C_{t-1} 通过或部分通过。然后进一步产生需要更新的新信息^[10]。需要更新的新信息包含两部分,第一部分是输入门层通过sigmoid函数决定哪些值用来更新,第二部分是tanh层用来生成新的候选值 C_{-t} ,它作为当前层产生的候选值会添加到记忆单元中。模型结合这两部分产生的值进行更新^[11]。

文中在识别中文电子病历中的命名实体时,将汉字分解成若干个偏旁部首,每个汉字表示成 d 维向量。对给定的包含 n 个汉字的句子 (x_1, x_2, \dots, x_n) ,句子中的每个汉字,LSTM模型通过式1~6计算字左边内容的 h_t 和字右边内容的 h_t ,得到词向量的LSTM表示,从而包含了所需要记忆的信息。

在LSTM神经网络模型中,直接用 h_t 作为特征值去计算网络输出 y_t ,在识别中文命名实体时,输出标签之间存在的一些约束条件LSTM模型无法表示出来。如文中所识别的五类中文命名实体,身体部位BOD、疾病名称DIS、检查手段EXA、症状SYM和治疗TRE,通常B表示开始的字,I表示中间的字,E表示最后的字,S表示该实体是单个字,I-BOD不能在B-DIS之后,LSTM模型无法表示这些约束条件,因此在LSTM模型中嵌入CRF模型,利用CRF模型计算输出 y_t 的值。

在条件随机场中,每个特征函数有下面几个输入值:一个句子 X 、一个单词在句子中的位置 i 、当前单词的标签 l_i 、前一个单词的标签 l_{i-1} 、输出为一个实数(通常是0或者1)^[12]。在LSTM-CRF模型中,首先定义了句子 X 输出标签序列 y 的分值 $s(X, y)$ 的计算公式。

$$s(X, y) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i} \quad (7)$$

其中, A 是转移矩阵,表示将所有状态一步转移的概率; P 是LSTM输出的矩阵, p_{i, y_i} 是假设从第 i 个字到第 j 个字作为一个实体的分值。根据 $s(X, y)$ 的值选择 y 。

输出 $y^* = \operatorname{argmax} s(X, y)$,其中 $y^* \in Y_x, Y_x$ 表示 y 所有可能的标签序列。所设计实现的LSTM-CRF模型结构如图1所示。

2 实验分析

2.1 实验环境

实验的硬件环境如下:处理器为intel@Corei7CPU

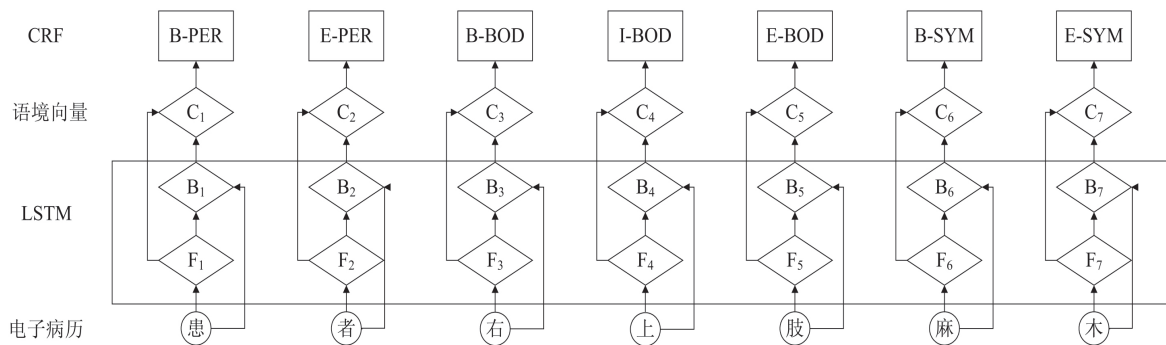


图 1 LSTM-CRF 模型

@ 3.60 GHz * 8; 内存 32 G; GPU 为 TITAN X (Pascal); 操作系统为 Ubuntu16.04。

文中设计实现的中文电子病历命名实体识别程序,使用 Python 程序设计语言开发,开发过程中调用的工具包如下:

jieba-0.38: 结巴分词模块可支持精确模式、全模式、搜索引擎模式三种分词方式,支持基于概率的用户词典。实验过程中使用精确模式并结合加载外部用户词典,从原文本产生分词语料。词典格式设计为一个词占一行,涵盖常用医学上的专有名词即确定的实体边界。

gensim-2.3.0 版本: gensim-2.3.0 是 Word2Vec 基于 python 的实现。Word2Vec 是 Google 公司发布的一个开源词向量工具包,并在语料中获取了高精度的词向量表示。实验中首先使用结巴分词库产生的分词语料来生成词向量,然后使用 Gensim 的 Word2Vec,训练结果在构建 LSTM 深度学习模型时使用。

tensorflow-gpu == 1.2.0 版本: 实现神经网络模型的开源工具。实验中使用 tensorflow-gpu == 1.2.0 搭建了 LSTM 神经网络模型^[13-14]。

2.2 实验数据处理

首先,在医学专家指导下人工标注了 100×4 条实体语料^[15],并建立字典,字典中包括实体和实体类型;其次,利用模型生成部分实体标注语料,并设计程序自动校对,校对程序判断模型生成的语料是否与字典中的一致,包括实体和实体类型是否一致;最后,生成深度学习模型需要的 BIOES 字标签形式语料。反复迭代下去,不断优化模型生成语料,直至建立好模型需要的语料。

2.3 实验结果

准确率是多分类中最重要的性能指标。该实验中的准确率达到 96.29%,精确率、召回率、F 值分别是 91.61%、96.22%、93.85%。所识别的 5 个实体的精确率、召回率和 F 值如表 1 所示。

表 1 中疾病名称和治疗的精确率相对较低,主要有两方面的原因。一是和训练测试数据不均衡相关,

因为电子病历中包含的相关信息相对较少;二是和词典相关,随着医学技术的发展,许多新的治疗方法并未录入词典中。

表 1 LSTM-CRF 多分类器性能评价指标

实体类别	精确率 / %	召回率 / %	F
身体部位	91.52	93.52	92.51
疾病名称	89.93	90.17	90.05
检查	90.52	98.81	94.48
症状	96.08	98.98	97.51
治疗	81.61	90.76	85.94

3 结束语

文中设计实现了基于 LSTM-CRF 的中文电子病历命名实体识别系统,该系统能识别五种实体类型,准确率达到了 96.29%,超过了大多数多分类识别器的准确率。实验结果为基于中文电子病历的关系抽取和构建医疗知识图谱奠定了扎实的基础。另外,该系统也存在不足,需要进一步改进,如基于 LSTM-CRF 模型的训练时间。实验中,在没有 GPU 的环境下训练,在人工标注的 400 条语料上,花费了 69 个小时,在 TITAN X (Pascal) GPU 的环境下训练,仍然花费了 3 个小时,因此下一步工作将进一步完善模型,以缩短训练时间。

参考文献:

- [1] 杨锦峰,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.
- [2] 中华人民共和国卫生部.电子病历系统功能应用水平分级评价方法及标准[EB/OL].2011-11-02.http://www.moh.gov.cn/mohyzs/s3586/201111/53274.shtml.
- [3] 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.
- [4] SUN Yaming, LIN Lei, TANG Duyu, et al. Radical-enhanced chinese character embedding[C]//21st international conference on neural information processing. Kuching, Ma-

(下转第 142 页)

试工具,主要用来测试 Web 应用,能够在浏览器中模拟用户操作。下面以淘宝为例,详细说明其使用方法。

(1) 分析淘宝主页面,通过在检索框中输入检索词,点击确定,就能得到相关商品信息;

(2) 构建基于 Scrapy 爬虫,改写下下载器中间件文件,在该文件中导入 Selenium 模块,创建 Selenium 的 Chrome 对象;

(3) 修改类中 process_request 方法,添加元素等待语句,一旦页面元素加载出来,就进行填入检索词和点击按钮两个操作;

(4) 运行以上爬虫,可以看到 Chrome 浏览器被自动打开,自动进行填入检索词和点击按钮两个操作,页面数据正常加载出来,爬虫就可以提取加载出来的数据。

4 结束语

爬虫作为获取数据的重要工具之一,广泛应用于各种网站。据统计,网络上百分之六十的流量都是爬虫产生的。文中仅仅列举了一些简单的爬虫和反爬虫方法,指出网站为保护数据所采取的措施,以及爬虫又是如何突破重重限制的。如今,爬虫和反爬虫之间的较量愈演愈烈,还有更复杂的爬虫和反爬技术值得去探讨。

参考文献:

[1] 闫立达,薛朋强.基于匿名网络的爬虫设计与实现

(上接第 108 页)

laysia: Springer, 2014: 681-687.

- [5] WANG Baoxun, LIU Bingquan, WANG Xiaolong, et al. Deep learning approaches to semantic relevance modeling for chinese question-answer pairs[J]. ACM Transactions on Asian Language Information Processing, 2011, 10(4): 21.
- [6] PENG Nanyun, DREDZE M. Named entity recognition for chinese social media with jointly trained embeddings[C]// Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon, Portugal: ACM, 2015: 548-554.
- [7] 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.
- [8] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]// Proceedings of the 2016 conference on human language technology. [s.l.]: ACM, 2016: 260-270.
- [9] LI Y, LI W, SUN F, et al. Component-enhanced chinese character embeddings[C]// Conference on empirical methods in natural language processing. [s.l.]: [s.n.], 2015: 829-

[J].现代计算机, 2017(16): 45-49.

- [2] 黄媛.面向网络爬虫的企业网站优化策略[J].信息系统工程, 2017(4): 23.
- [3] 安子建.基于 Scrapy 框架的网络爬虫实现与数据抓取分析[D].长春: 吉林大学, 2017.
- [4] 魏冬梅,何忠秀,唐建梅.基于 Python 的 Web 信息获取方法研究[J].软件导刊, 2018, 17(1): 41-43.
- [5] 刘宇,郑成焕.基于 Scrapy 的深层网络爬虫研究[J].软件, 2017, 38(7): 111-114.
- [6] TAN Qingzhao, MITRA P. Clustering-based incremental web crawling[J]. ACM Transactions on Information Systems, 2010, 28(4): 85-89.
- [7] HTTP threats evade normal protections[J]. Computer Fraud & Security, 2013, 2013(9): 132-136.
- [8] 董博,李翀,刘学敏,等.基于爬虫的数据监控系统[J].计算机系统应用, 2017, 26(10): 53-60.
- [9] 赵本本,殷旭东,王伟.基于 Scrapy 的 GitHub 数据爬虫[J].电子技术与软件工程, 2016(6): 199-202.
- [10] ZHENG Qinghua, WU Zhaohui, CHENG Xiaocheng, et al. Learning to crawl deep web[J]. Information Systems, 2013, 38(6): 801-819.
- [11] 马联帅.基于 Scrapy 的分布式网络新闻抓取系统设计与实现[D].西安: 西安电子科技大学, 2015.
- [12] 李代祯,谢丽艳,钱慎一,等.基于 Scrapy 的分布式爬虫系统的设计与实现[J].湖北民族学院学报: 自然科学版, 2017, 35(3): 317-322.
- [13] 洪芳.基于 Selenium2 的 Web UI 自动化测试框架的设计与实现[D].成都: 西南交通大学, 2017.
- [14] MA Xueze, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]// Proceedings of the 54th annual meeting of the association for computational linguistics. [s.l.]: ACM, 2016: 125-132.
- [11] 孙志军,薛磊,许阳明,等.深度学习研究综述[J].计算机应用研究, 2012, 29(8): 2806-2810.
- [12] CHIU J P C, NICHOLS E. Named entity recognition with bi-directional LSTM-CNNs[J]. Transactions of the ACL, 2015, 4(1): 357-370.
- [13] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures[C]// Proceedings of The 32nd international conference on machine learning. Lille, France: [s.n.], 2015: 2342-2350.
- [14] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space[C]// Proceedings of the international conference on learning representations. Roston, VA, USA: Internet Society, 2013: 1-12.
- [15] 蒋志鹏,赵芳芳,关毅,等.面向中文电子病历的词语标注研究[J].高技术通讯, 2014, 24(6): 609-615.