计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(2):8−14 [doi: 10.15888/j.cnki.csa.006760]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

在线医疗问答文本的命名实体识别①

杨文明, 褚伟杰

(北京大学 软件与微电子学院, 北京 102600)
通讯作者: 褚伟杰, E-mail: chuwj@ss.pku.edu.cn

摘　要: 本文主要是对在线问诊中产生的医疗文本进行命名实体识别的研究. 使用在线医疗问答网站的数据, 采用{B, I, O}标注体系构建数据集, 抽取疾病、治疗、检查和症状四个医疗实体. 以 BiLSTM-CRF 为基准模型, 提出两种深度学习模型 IndRNN-CRF 和 IDCNN-BiLSTM-CRF, 并在自构建数据集上验证模型的有效性. 将新提出的两种模型与基准模型通过实验对比得出: 模型 IDCNN-BiLSTM-CRF 的 F1 值 0.8116, 超过了 BiLSTM-CRF 的 F1 值 0.8009, IDCNN-BiLSTM-CRF 整体性能好于 BiLSTM-CRF 模型; 模型 IndRNN-CRF 的精确率 0.8427, 但该模型在召回率上低于基准模型 BiLSTM-CRF.

关键词: 医疗问答; 深度学习; 独立循环神经网络; 膨胀卷积; 双向循环神经网络

引用格式: 杨文明,褚伟杰.在线医疗问答文本的命名实体识别.计算机系统应用,2019,28(2):8–14. http://www.c-s-a.org.cn/1003-3254/6760.html

Named Entity Recognition of Online Medical Question Answering Text

YANG Wen-Ming, CHU Wei-Jie

(School of Software & Microelectronics, Peking University, Beijing 102600, China)

Abstract: This paper mainly presents the research of named entity recognition of medical texts generated by online inquiry. Using the data of online medical quiz website, we employ {B, I, O} annotation system to build data sets, and extract four medical entities of disease, treatment, examination, and symptom. Taking BiLSTM-CRF as the benchmark model, two deep learning models IndRNN-CRF and IDCNN-BiLSTM-CRF are proposed, and the validity of the model on the self built dataset is verified. The two new models are compared with the benchmark model by experiment. It is concluded that the model IDCNN-BiLSTM-CRF has an F1 value of 0.8165, which exceeds the BiLSTM-CRF's F1 value of 0.8009. The overall performance of IDCNN-BiLSTM-CRF is better than that of BiLSTM-CRF. The IndRNN-CRF model has a high precision rate of 0.8427, but its recall rate is lower than the benchmark model BiLSTM-CRF.

Key words: medical question and answer; deep learning; Independent Recurrent Neural Network (IndRNN); dilation convolution; bi-directional RNN

1　引言

伴随互联网和大数据技术的发展, 很多患者在感到身体不适时, 首先会到医疗问答网站上提问和查询疾病相关的问题, 同时许多医生也会到医疗问答网站去回答患者的疑问, 这类网站已经成为联系患者和医生之间的枢纽. 在线医疗问答社区的发展使得我们获取医学知识的渠道多样化, 有助于患者了解自己的健康状况, 同时也有助于健康医学知识的普及. 国内如 39 健康网, 寻医问药, 春雨医生等网站, 不仅提供基础的疾病知识和医学知识, 而且每天还积累了大量的问答数据, 这些医疗文本数据中包含大量有意义的信息, 如寻医问药从 2004 年开始一直到现在, 已经积累了大

① 收稿时间: 2018-07-31; 修改时间: 2018-08-30; 采用时间: 2018-09-11; csa 在线出版时间: 2019-01-28

量真实的信息, 并且每天都在产生数万条的问答数据. 这些医疗文本数据将汇聚成非常客观的大数据, 数据中包含有大量的真实案例和医生的诊疗建议. 在这些数据中蕴含着比较丰富的医疗价值. 但这些数据是非结构化的状态, 无法进行更深的数据挖掘, 实现数据的充分利用. 为了更好的利用这类数据, 抽取和挖掘更有价值的医疗信息, 需要把非结构化的数据进行结构化, 而命名实体识别是结构化文本中的第一步, 而且该类文本的命名实体识别可以为医疗问答的研究和应用打好基础. 本文利用医疗问答网站积累的数据, 进行了命名实体识别的研究.

2 命名实体识别相关研究

实体是文本的基本信息元素, 是构成文本的基础. 命名实体识别 (Named Entity Recognition, NER) 是自然语言处理的一项基本任务, 主要是从一段文本中找出实体, 并对实体出现的位置和类别进行标记. NER 概念的提出是在 MUC-6 (Message Understanding Conference) 会议上[1], 最初的提出是作为信息提取的重要任务之一. 通用的命名实体识别任务, 主要是在一段文本中识别出人名, 地名, 专业机构, 时间和数字 (货币, 百分数) 等. 在特定的领域, 可以用来识别特殊领域的实体如医疗领域和金融领域等. 命名实体识别技术包括许多不同的方法: 基于词典和规则的方法; 基于统计学习的方法; 还有将二者混合的方法. 常见的统计学习的方法有支持向量机 (SVM)、最大熵模型、贝叶斯分类等, 这些方法把 NER 任务看成分类问题. 此外, 还有隐马尔可夫模型 (HMM) 和条件随机场 (CRF), 这类模型把 NER 任务当做序列标注问题处理. 随着深度学习的发展, 最近几年出现了大量的基于神经网络的模型, 并取得了较好的效果, 最具代表性的是 BiLSTM-CRF 模型[2], 该模型在各个公共数据集上均取得了不错的效果. 在 RNN 的输出层连接 CRF 层, 这种结构已经成为命名实体识别模型的常用结构. 目前对于医疗文本命名实体识别的研究主要集中在电子病历, 医学文献, 医学书籍等, 而互联网医疗问答社区文本的研究并不多, 国内最近几年也有研究者开始关注这方面的研究, 比如苏娅等[3]使用 CRF 在自建数据集上进行研究, 抽取的目标实体共 5 类, 分别包括疾病、症状、药品、治疗方法和检查, 通过采用逐一添加特征的方式训练模型, 模型精确率达到 81.26%, 召回率 60.18%. 张帆等人[4]设计深度神经网络应用到在线医疗文本实体识别上, 抽取的目标实体也是 5 类. 深度神经网络模型同 CRF 等方法相比减少了很多人工特征, 并且提高了精确率和召回率.

3 算法模型设计

本文以 BiLSTM-CRF 作为基准模型, 设计了两种不同的命名实体识别模型, 并且在自构建的数据集上进行验证, 均取得了不错的效果.

3.1 BiLSTM-CRF 模型

双向循环神经网络 (BiLSTM) 由两个单向的循环神经网络构成, 两个网络中一个随时间正向, 另一个随时间逆向, 逆向网络的实现本质上把输入序列进行逆转, 然后输入到正向网络中. BiLSTM 的优势是可以在当前节点获取正反两个方向的特征信息, 即能捕捉到未来信息的特征, 也能捕捉到过去信息的特征. 但是, 两个方向的循环神经网络并不会共享一个隐状态, 正向 LSTM 的隐状态传给正向的 LSTM, 逆向 LSTM 的隐状态传给逆向的 LSTM, 两个方向的循环神经网络之间没有连接, 两个输出会共同连接到输出节点合成最终输出. 方向不同的两个循环神经网络, 都可以展开成为普通的前馈网络, 使用反向传播算法 (BPTT) 进行训练. 双向循环神经网络被用在许多序列标注任务上.

条件随机场 (CRF) 模型是一种概率无向图模型, 可以解决序列标注任务, 命名实体识别可以看做是序列标注任务, 即给定观察序列 $X=\{x_1, x_2, \cdots, x_n\}$的条件下, 求 Y 的概率. 随机变量 $Y=\{y_1, y_2, \cdots, y_n\}$, Y 是隐状态序列. 数学表达式为 $P(Y|X)$. 在命名实体识别上使用的 CRF 主要是 CRF 线性链, CRF 建模的数学公式如式 (1) 和 (2).

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{k=1}^{K} w_k f_k(y,x)\right) \tag{1}$$

$$Z(x)=\sum_{y}\exp\sum_{k=1}^{K} w_k f_k(y,x) \tag{2}$$

上式中 f_k 是特征函数, w_k 是特征函数的权重, $Z(x)$ 是归一化因子. 条件随机场可以看成是定义在序列上的对数线性模型, 能够使用极大似然估计方法求参数. 目前已经有了一些优化算法进行该问题的求解, 比如梯度下降法, 改进的迭代尺度法和拟牛顿法等. 模型在进行解码时可以利用维特比算法, 这是一种动态规划算法,

在给定观察序列的条件下, 求出最大的标记序列的概率. BiLSTM 与 CRF 的结合, 本质上是把 BiLSTM 的输出作为 CRF 的输入, BiLSTM 层输出的是每一个标签的预测分值, 这些分值会输入到 CRF 层. 其过程可描述为利用 BiLSTM 解决提取序列特征, 再使用 CRF 利用句子级别的标记信息进行训练, 单独使用 BiLSTM 也可以完成命名实体识别, 可以从 BiLSTM 的输出中挑选最大值对应的标签, 作为该单元的标签, 但是这不能保证每次预测的标签都是合法的, 比如对于{B, I, O}体系的标注, 标签序列是"I-Organization I-Person"和"B-Organization I-Person", 很显然这是错误的. 如果在 BiLSTM 的输出层接入 CRF 层后, 相当于对最后的预测标签加入了约束, 保证输出的标签是合法的, 这些约束会在训练的过程学习到, 对于 BiLSTM-CRF 模型的学习方法同样可以使用极大似然估计方法.

可以把双向循环神经网络的输出看成打分矩阵, 称为 P 矩阵. 对于输入语句 $X=(x_1, x_2, x_3, \cdots, x_n)$, P 是一个 $n\times k$ 的矩阵, k 是输出标注 y 的个数, $P_{i,j}$ 表示句子中第 i 个词被标记为第 j 个标签的概率. 句子的预测标注序列可以表示为: $y=(y_1, y_2, y_3, \cdots, y_n)$. 定义 y 矩阵的打分函数的计算式 (3).

$$score(X,y)=\sum\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{3}$$

$A_{i,j}$ 是看成转移打分矩阵, 代表从标注 i 转移到标注 j 的得分. y_0 和 y_n 分别代表句子开始和结束的标签, 标注矩阵 A 是一个 $k+2$ 阶的方阵. 通过式 (4) 计算 y 在给定 x 下的条件概率 $p(y|x)$, 其中 Y_X 代表对于给定的句子 X 所有可能的标签序列, 损失函数可以定义为式 (5), 并在训练的过程中极大化正确标签序列概率的对数值.

$$P(y|X)=\frac{\exp(score(X,y))}{\sum_{\tilde{y}\in YX}\exp(X,\tilde{y})} \tag{4}$$

$$L=\log(P(y|X)) \tag{5}$$

在模型训练完成后可以通过式 (6) 进行模型预测, 其中 $y*$是集合中使得得分函数 $score$ 取最大值的序列.

$$y*=\arg\max_{\tilde{y}\in YX}score(X,\tilde{y}) \tag{6}$$

以上是 BiLSTM-CRF 的基本原理, 本文在设计的 BiLSTM-CRF 模型结构图如图 1. 首先将输入语句经过一个 embedding 层, 之后连接到 BiLSTM 层, 在 BiLSTM 层后连接映射层, 并进行逻辑回归, 该层的输出会输入到下一层 CRF 层. 为了提高模型的泛化能力, 在 embedding 层和 BiLSTM 层之间加入了 droupout 层.

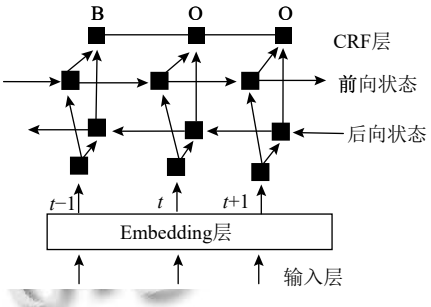

图 1 BiLSTM-CRF 模型结构图

3.2 IndRNN-CRF 模型

独立循环神经网络 (Independently Recurrent Neural Network, IndRNN)[5]由 Li S 等人提出, 同传统的 RNN 不同的是 IndRNN 的神经元之间是独立的. 传统的 RNN 的隐藏层数学公式如式 (7), 其中 W 是 $N\times M$ 的矩阵, U 是 $N\times N$ 的矩阵, N 是 RNN 中的神经元节点数.

$$h_t=\sigma(Wx_t+Uh_t-1+b) \tag{7}$$

在传统的 RNN 中每个神经元都和上一时刻的全部神经元发生联系 (U 的行向量与 h_{t-1} 向量的乘积, h_{t-1} 是 t–1 时刻的隐状态), 也就是神经元之间是不独立的. 而 IndRNN 结构神经元之间的连接仅发生在层与层之间, IndRNN 的数学表达式可以在上面式 (7) 进行改造后得到如式 (8). 其中 U 和 h_{t-1} 是点积, 此时的 U 不是矩阵, 而是一个 N 维的向量, t 时刻的每个神经元只和 t–1 时刻自身相联系, 与其他的神经元无关. 这也是独立循环神经网络名称的由来. 为了在神经元之间发生联系, 至少需要进行两层的堆叠.

$$h_t=\sigma(Wx_t+U\otimes h_{t-1}+b) \tag{8}$$

模型中的第 n 个神经元的隐藏状态 $h_{n,t}$, 可由式 (9) 计算得出, 其中 w_n 和 u_n 分别是输入权重和 t–1 到 t 时刻的连接权重的第 n 行, 每个神经元只接受前一步它自己的隐藏状态和输入传来的信息. 这与传统的 RNN 是不同的, 这种结构提供了一种循环神经网络的新视角, 随着时间的推移 (通过 u), 独立的聚集空间模式 (通过 w), 不同神经元的相关性可以通过多层堆叠来实现, 下一层的神经元处理上一层所有神经元的输

出. 模型同样采用梯度后向传播算法进行优化, IndRNN 进一步缓解了随时间累积的梯度爆炸或消失的问题, 梯度可以在不同的时间步上有效的传播, 可以使得网络叠加更深.

$$h_{n,t} = \sigma(w_n x_t + u_n h_{n,t-1} + b_n) \tag{9}$$

在本文中, 将 BiLSTM-CRF 模型中的 BiLSTM 换成多层的 IndRNN, 提出了一种新的模型 Multi-IndRNN-CRF, 本文中 IndRNN 有 4 层, 之后拼接 CRF, 模型示意图如图 2, 在 Embedding 层后输出的数据, 会进行 dropout, 每层 IndRNN 输出后都会进行 BatchNormalization 防止数据发生严重偏移, 同时防止梯度爆炸. IndRNN-CRF 的损失函数和 LSTM-CRF 模型一样, 参数的学习方法依然是极大似然估计.

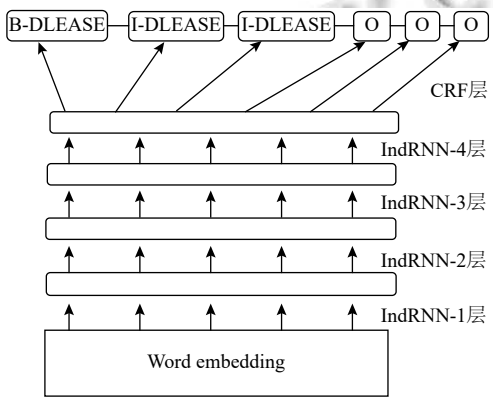

图 2 4-IndRNN-CRF 结构图

3.3 IDCNN-BiLSTM-CRF 模型

膨胀卷积 (Dilated Convolution, 简称 DCNN) 是 Yu F, Koltun V 在 2015 年提出的[6], 经典卷积的 filter, 作用在矩阵的一片连续区域上做滑动, 而膨胀卷积是在 filter 中增加了膨胀宽度, 在输入矩阵上做滑动时会跳过膨胀宽度中间的数据. filer 矩阵大小不变, filter 最终获取到了更广的输入矩阵的数据. DCNN 的示意图如图 3.

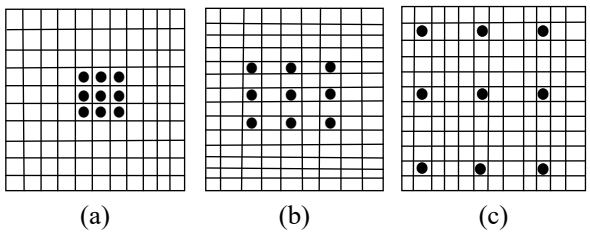

图 3 DCNN 示意图

图 3 中的 (a) 图对应 3×3 的 1-dilated convolution, 同经典的卷积操作一样; (b) 图对应 3×3 的 2-dilated convolution, 卷积核的大小仍然是 3×3, 空洞大小是 1, 可以理解成卷积核的大小是 7×7, receptive filed 是 7×7; (c) 图是 4-dilated convolution 操作, receptive filed 是 15×15 的感受野.

Strubell E 等人[7]提出了 IDCNN(Iterated Dilated Convolution, IDCNN) 模型, 用在实体识别任务上取得了不错的效果. 膨胀的宽度随着层数的增加呈现为指数增加, 但参数的数量是线性增加的, 这样接受域很快就覆盖到了全部的输入数据. 模型是 4 个大小相同的膨胀卷积块叠加在一起, 每个膨胀卷积块里的膨胀宽度分别为 1, 1, 2 的三层膨胀卷积. 把句子输入到 IDCNN 模型中, 经过卷积层, 提取特征, 其基本框架同 BiLSTM-CER 一样, 由 IDCNN 模型的输出经过映射层连接到 CRF 层.

尽管 IDCNN 模型可以使得接受域扩大, 但不会像双向循环神经网络, 可以从序列的整体提取正向和反向特征, 但循环神经网络不能很好的兼顾到局部特征, 本文提出模型 IDCNN-BiLSTM-CRF 既能兼顾全局特征 (通过 BiLSTM), 又能兼顾局部特征 (通过 IDCNN).

模型的基本结构描述如下: 首先输入语句经过 embedding 层, 输出字向量, 字向量并行输入到 IDCNN 模型和 BiLSTM 模型, 经过两个模型后, 将输出的向量进行拼接后形成向量特征, 然后经过映射层后输入到 CRF 层. IDCNN 在提取局部特征的同时能够兼顾到部分全局特征, 但不会像 BiLSTM 能够很好的提取全局特征, 因此将 IDCNN 输出的向量特征作为对局部特征的弥补, 拼接在 BiLSTM 的向量特征上, 模型示意图如图 4. 其中 Dilated CNN block 中有三个卷积层, 没有池化层, 第一层为 1-dilated convolution, 第二层为 1-dilated convolution, 第三层为 2-dilated convolution. 将 4 个 block(对应图中的 DCNN-i) 堆叠, 当有数据输入后. 先经过 embedding 层, 然后输入到 DCNN-1, 从 DCNN-1 的输出有两个去向, 一是与其他 DCNN 的输出拼接后形成最终的特征向量, 另一个输出变成 DCNN-2 的输入, 依次类推. IDCNN 模型的输出和 BiLSTM 模型的输出进行拼接后, 经过一个映射层, 再将值输入到 CRF 中, 模型图见图 4. 模型的训练以及参数的学习方法同 BiLSTM-CRF.

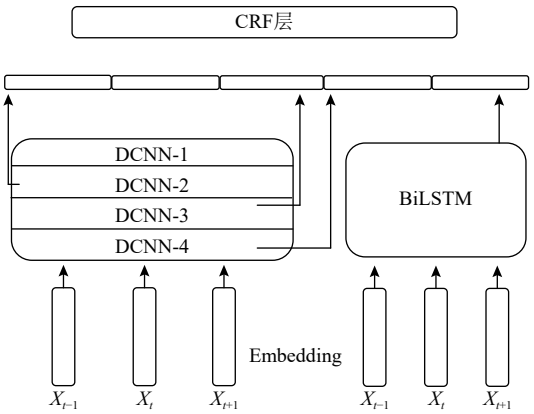

图 4 IDCNN-BiLSTM-CRF 模型结构图

4 数据处理和标注

使用 Scrapy 框架编写爬虫, 从医疗问答网站爬取数据. 爬取的网站分别是“寻医问药”, “39 健康网”, “快速问医生”等网站. 在各个网站的问答板块收集咨询者的问题, 总计收集数据约 1200 万条, 大约 2.27 GB. 从收集的数据中挑选 8027 条数据作为训练集, 1972 条作为测试集. 采用{B, I, O}标注体系, 对医疗文本进行人工标注, 具体格式为 B-X, I-X 和 O. B 代表实体开始, I 代表实体中间或结束部分, O 代表非实体. 标注的实体类别参考杨锦峰等人[8]在论文中提出的方案, 分为四类实体: 疾病、症状、检查和治疗. X 代表命名实体的类别, 分别为 DISEASE、SYMPTOM、TREATMENT、CHECK 四个不同的标识, 代表疾病、症状、治疗和检查. 该任务的标记一共有 9 (=4×2+1) 类标签, 在标注时把药物归结到治疗实体. 表 1 给出了标注的示例.

表 1 命名实体类别

标识符号	实体类别	类别的定义	示例
B-DISEASE I-DISEASE	疾病	主要对应于 ICD-10(国际疾病分类) 中定义的疾病术语	肺 B-DISEASE 癌 I-DISEASE
B-SYMPTOM I-SYMPTOM	症状	由于疾病引起的不适的表现或异常的表现, 可以含有身体部位	发 B-SYMPTOM 烧 I-SYMPTOM
B-TREATMENT I-TREATMENT	治疗	给患者提供的干预措施, 药物, 手术	青 B-TREATMENT 霉 I-TREATMENT 素 I-TREATMENT
B-CHECK I-CHECK	检查	为了证明疾病而采取的检测方法	胸 B-CHECK 片 I-CHECK
O	非实体	凡不属于实体的字符或汉字	, O 否 O

实际标注的格式如下: 对于语句“我有点发烧, 浑身无力, 是感冒了吗？”, 标注为{O, O, O, B-SYMPTOM, I-SYMPTOM, O, B-SYMPTOM, I-SYMPTOM, I-SYMPTOM, I-SYMPTOM, O, O, B-DISEASE, I-DISEASE, O, O, O}, 其中语句中包含的标点符号作为非实体, 标注为“O”.

5 实验结果和分析

5.1 实验条件

本文实验是在 Linux 平台下使用 Python 3.5 语言在 tensorflow 框架进行开发, 硬件环境如下: Intel i7 的 cpu, 16 GB 内存以及 NVIDIA GTX-1070 显卡.

使用预训练的字向量对 embedding 层进行初始化, 预训练字向量的过程如下: 首先将下载的 1200 万条问句, 按照字级别进行字向量的训练. 训练的模型使用的是开源工具 Word2vec, 该工具是 Toms Mikolov 在 2013 年开发的工具包, Word2vec 使用 CBOW 模型[9–11] (连续词袋模型). 对于 word2vec 参数的设定如下: 字向量的维度设置为 200, 窗口大小为 5, 训练次数为 20, 其余参数默认.

5.2 模型参数设置

对于 BiLSTM-CRF 模型参数的设定: BiLSTM 的隐层节点为 300, 模型中的 droupout 层参数设置为 0.5, 采用 Adam 优化算法, 学习率设置为 0.001, batch size 的大小为 64, epoch 的大小为 100.

对于 IndRNN-CRF 模型参数的设定: IndRNN 的隐层节点为 300, 共有 4 层 IndRNN, 模型中的 droupout 层参数设置为 0.5, 采用 Adam 优化算法, 学习率设置为 0.001, batch size 的大小为 64, epoch 的大小为 100.

对于 IDCNN-BiLSTM-CRF 模型参数的设定: BiLSTM 的隐层节点为 300, IDCNN 的 filter 个数为 100, 模型中的 droupout 层参数设置为 0.5, 采用

Adam 优化算法, 学习率设置为 0.001, batch size 的大小为 64, epoch 的大小为 80.

5.3 实验结果和分析

实验结果的评价指标有 3 个, 分别为精确率, 召回率和 F 值. 计算公式如式 (10), (11), (12).

$$Precision(P)=\frac{\text{系统正确识别的实体个数}}{\text{系统识别的实体个数}} \tag{10}$$

$$Recall(R)=\frac{\text{系统正确识别的实体个数}}{\text{文档中的实体个数}} \tag{11}$$

$$F-measure=\frac{2\times P\times R}{P+R} \tag{12}$$

不同模型的实验结果分别见表 2, 3 和 4.

表 2 BiLSTM-CRF 的实验结果

实体类别	精确率	召回率	F1 值
SYMPTOM	0.8243	0.8145	0.8194
DISEASE	0.9458	0.9172	0.9313
TREATMENT	0.7727	0.7961	0.7843
CHECK	0.7171	0.6267	0.6689
平均值	0.8150	0.7886	0.8009

对比实验结果可以看出, IndRNN-CRF 模型在精确率上比基准模型 BiLSTM-CRF 高, 召回率的值为 0.6848, 相比于模型 BiLSTM-CRF 的召回率比较低. IDCNN-BiLSTM-CRF 模型在精确率, 召回率和 F1 值上均超过了基准模型 BiLSTM-CRF. 图 5, 图 6 和图 7 分别是模型 BiLSTM-CRF, IndRNN-CRF 和 IDCNN-BiLSTM-CRF 的 *Loss* 曲线图, 纵坐标代表 *Loss* 值, 横坐标代表的是迭代次数. 从图中可以看出在经过了 24 000 次的迭代后模型 BiLSTM-CRF 的 *Loss* 值大于 2.0, 模型 IndRNN-CRF 和 IDCNN-BiLSTM-CRF 的 loss 值小于 2.0, 其中模型 IndRNN-CRF 的 loss 值最低.

表 3 IndRNN-CRF 的实验结果

实体类别	精确率	召回率	F1 值
SYMPTOM	0.8640	0.7171	0.7837
DISEASE	0.9492	0.9066	0.9274
TREATMENT	0.7935	0.6226	0.6977
CHECK	0.7643	0.4931	0.5994
平均值	0.8427	0.6848	0.7521

表 4 IDCNN-BILSTM-CRF 的实验结果

实体类别	精确率	召回率	F1 值
SYMPTOM	0.8669	0.8023	0.8334
DISEASE	0.9469	0.9172	0.9318
TREATMENT	0.8076	0.8044	0.8060
CHECK	0.7727	0.6313	0.6949
平均值	0.8485	0.7888	0.8165

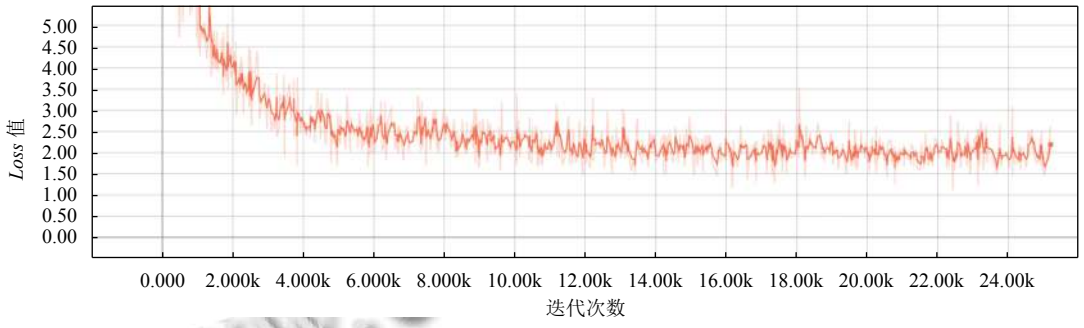

图 5 BiLSTM-CRF 的 loss-step 曲线图

由于 IDCNN-BiLSTM-CRF 模型的总体性能最好, 可以用在互联网在线问诊医疗文本的实体识别上, 该模型也可用在医学文献, 电子病历等文本的命名实体识别上. 模型 IndRNN 可以用在对精确率要求较高, 但对召回率要求不高的任务中.

6 结论与展望

本文针对在线问诊医疗文本, 利用深度学习技术设计了两种不同的神经网络模型, 进行医疗文本命名实体识别的研究, 共识别 4 类医疗实体: 疾病, 症状, 治疗和检查. 对基于字级别的命名实体识别任务, 在模型 IDCNN-BiLSTM-CRF 中使用卷积神经网络和循环神经网络提取特征向量, 并将两个特征向量拼接, 形成既包含全局特征又包含局部特征的向量, 该向量经过映射层后输入到 CRF 层中, 实验结果表明该模型的整体性能最好. 但是由于医疗领域的特殊性, 仍然需要继续提高医疗实体的识别率, 获取更精确的挖掘结果. 在接下来的工作中, 可以考虑先对医疗文本分词, 然后加入

词性或者拼音等特征训练模型, 提高识别率. 此外, 对于医疗文本还要考虑文本中是否含有修饰性实体, 比如表示时间和否定的词汇等, 如“无头痛”, 症状“头痛”前的“无”就是修饰实体. 模型最终结果与参数的调试也有较大的关系, 设置不同的参数, 模型的输出值可能会不同.

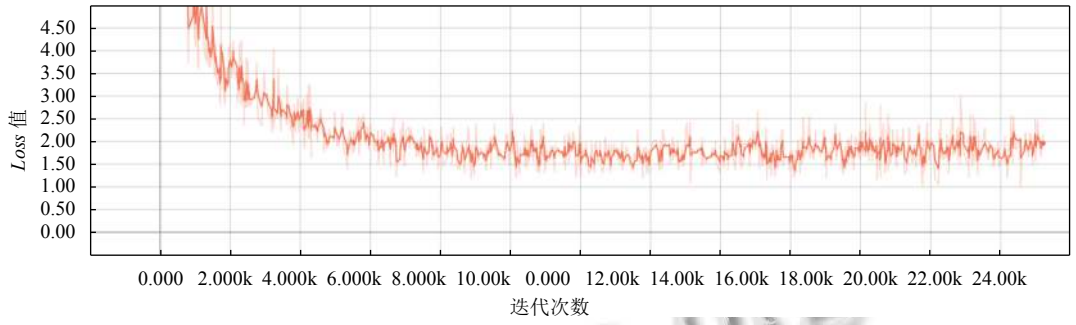

图 6 IndRNN-CRF 的 loss-step 曲线图

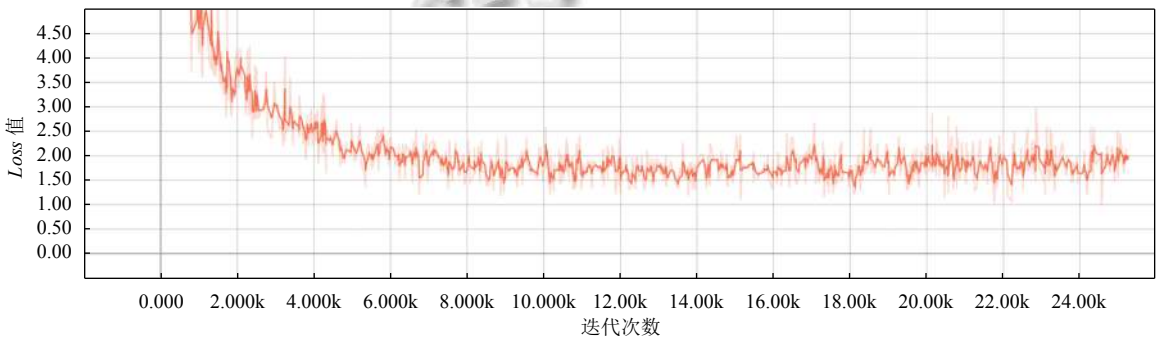

图 7 IDCNN-BILSTM-CRF 的 loss-step 曲线

参考文献

1 Sundheim BM. Named entity task definition, version 2.1. Proceedings of the Sixth Message Understanding Conference. Columbia, MA, USA. 1995. 319–332.

2 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv: 1508.01991, 2015.

3 苏娅, 刘杰, 黄亚楼. 在线医疗文本中的实体识别研究. 北京大学学报 (自然科学版), 2016, 52(1): 1–9.

4 张帆, 王敏. 基于深度学习的医疗命名实体识别. 计算技术与自动化, 2017, 36(1): 123–127. [doi: 10.3969/j.issn.1003-6199.2017.01.025]

5 Li S, Li WQ, Cook C, *et al*. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN. arXiv preprint arXiv: 1803.04831, 2018.

6 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv: 1511.07122, 2016.

7 Strubell E, Verga P, Belanger D, *et al*. Fast and accurate entity recognition with iterated dilated convolutions. Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark. 2017. 2670–2680.

8 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. 自动化学报, 2014, 40(8): 1537–1562.

9 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. ICLR: Proceeding of the International Conference on Learning Representations Workshop Track. AZ, USA. 2013. 1301–3781.

10 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 3111–3119.

11 Kenter T, Borisov A, de Rijke M. Siamese CBOW: Optimizing word embeddings for sentence representations. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. 2016. 941–951.