计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):375−381 [doi: 10.15888/j.cnki.csa.008723]

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

融合文字与标签的电子病历命名实体识别①

赵　奎[1,2], 杜昕娉[1,2], 高延军[3], 马慧敏[4]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学, 北京 100049)
[3](中国医科大学附属第四医院, 沈阳 110032)
[4](东软集团股份有限公司 医疗解决方案事业本部, 沈阳 110003)
通信作者: 杜昕娉, E-mail: kexi_i@163.com

摘　要: 准确的命名实体识别是结构化电子病历的基础, 对于电子病历规范化编写有着重要的作用, 而现今的分词工具对于专业的医疗术语无法做到完全正确的区分, 使得结构化电子病历难以实现. 针对医疗实体识别中出现的问题, 本文提出了一种在命名实体识别领域中改进的 BiLSTM-CRF 深度学习模型. 模型将文字和标签结合作为输入, 在多头注意力机制中使模型关注更多的有用信息, BiLSTM 对输入进行特征提取, 得到每个文字在所有标签上的概率, CRF 在训练过程中学习到数据集中的约束, 进行解码时可以提高结果的准确率. 实验使用人工标注的 1 000 份电子病历作为数据集, 使用 BIO 标注方式. 从测试集的结果来看, 相对于传统的 BiLSTM-CRF 模型, 该模型在实体类别上的 *F*1 值提升了 3%–11%, 验证了该模型在医疗命名实体识别中的有效性.

关键词: 结构化电子病历; 命名实体识别; BiLSTM; CRF; 深度学习

引用格式: 赵奎,杜昕娉,高延军,马慧敏.融合文字与标签的电子病历命名实体识别.计算机系统应用,2022,31(10):375–381. http://www.c-s-a.org.cn/1003-3254/8723.html

Named Entity Recognition of Electronic Medical Records Based on Texts and Labels

ZHAO Kui[1,2], DU Xin-Ping[1,2], GAO Yan-Jun[3], MA Hui-Min[4]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](The Fourth Affiliated Hospital of China Medical University, Shenyang 110032, China)
[4](Medical Solutions Business Division, Neusoft Group Co. Ltd., Shenyang 110003, China)

Abstract: Accurate named entity recognition is the basis of structured electronic medical records and plays an important role in the standardized writing of electronic medical records. However, current word segmentation tools cannot completely and correctly distinguish professional medical terms, making it difficult to achieve structured electronic medical records. As for problems in medical entity recognition, this study proposes an improved deep learning model based on BiLSTM-CRF in the field of named entity recognition. The model combines text and labels as input, which makes the model focus on more useful information in the multi-head attention mechanism. BiLSTM performs feature extraction on the input and obtains the probability of each text on all labels. CRF learns the constraints of the data set during the training and improves the accuracy of the results after decoding. The experiment uses 1 000 manually labeled electronic copies as the data set and the BIO for labeling. Compared with the traditional BiLSTM-CRF model, the proposed model raises the *F*1 value in the entity category by 3%–11%, verifying its effectiveness in named entity recognition of medical records.

Key words: structured electronic medical record; named entity recognition; BiLSTM; CRF; deep learning

① 基金项目: 国家水体污染控制与治理科技重大专项 (2018ZX07601001)
收稿时间: 2022-01-10; 修改时间: 2022-01-30; 采用时间: 2022-02-22; csa 在线出版时间: 2022-06-28

1 引言

医疗电子病历[1]现今在临床诊断中被广泛应用, 近几年伴随着机器学习、深度学习的快速发展, 结构化电子病历得到了大众的关注. 由于难以将医生对病症的描述进行统一, 使得结构化医疗术语无法建立. 具体来说, 对于同一种疾病, 不同的医生在表达方式、中文的繁简体, 英文字母的大小写上的区别, 导致在医疗领域难以形成规范化的标准.

命名实体识别 (NER) 是自然语言处理 (NLP) 中信息抽取任务的一种, 在 NLP 的基础建设中有着较为广泛的应用. 结构化的电子病历需要实体的准确描述, 不能存在歧义或表述不明. 中文由于没有明确的分隔符, 使得实体识别的难度大大增加. 就模型训练而言, 使用词作为最小粒度还是字作为最小粒度会随着不同的应用场景有不同的效果, 难以确定使用哪种方式最好.

就提高命名实体识别的准确率方面, 孟捷[2]使用条件随机场 CRF 先对文本进行分词, 之后对分词结果进行属性标注, 并在词典中引入 ICD-10 标准, 使得实体识别取得了较好的效果. 江涛[3]提出了一种兼顾字词并通过自注意力机制延长实体联系距离的 WC-LSTM 模型, 使用 Word2Vec 训练的 100 维词向量嵌入模型, 并与 Bi-LSTM 和 Lattice-LSTM 模型进行对比. 沈宙锋等人[4]基于 XLNet-BiLSTM 模型, 通过对电子病历的序列化表示, 使得一词多义的问题得到了更好的解决. 王若佳等人[5]在使用无监督学习的 AC 自动机上对中文电子病历进行分词, 结合条件随机场和不同的实体类别, 得到了较好的识别模板. Zhang 等人[6]为解决中文实体的边界问题, 提出了一种基于单词方法和基于字符方法进行识别的方式, 在研究中, 为解决字符缺乏词级别信息和词边界信息, 使用了一种融合自匹配词特征的神经实体识别模型, 并在训练集上取得了不错的效果. Ji 等人[7]提出一种句子级的基于多神经网络模型的协同协作方法来进行实体识别, 通过 Word2Vec、GloVe、ELMo 对特定汉字嵌入预训练, 在 BiLSTM-CRF 和 CNN 模型上进行了模型测试. 李丹等人[8]提出了一种部首感知的识别方法, 该方法将部首信息编码到字向量中, 利用 BiLSTM-CRF 结合 Bert 模型, 使实体识别有一定的提高.

在过去的研究中, 大部分研究人员着重在于对文字进行处理, 而对标签在文本的上下文中的作用以及文本的预处理结果是否符合规范缺少关注. 查阅资料中考虑到, 如果可以在研究文字的同时挖掘电子病历标签中的隐含信息, 对于命名实体识别工作会有帮助. 例如, 病历中常用“宫颈癌”作为疾病术语, 那么在宫颈一词中被预测为疾病的实体, 应该以 3 个字的可能性较大, 不会出现诸如“宫颈的癌”或“宫颈的癌变”这样的实体, 从而提高实体识别的准确性, 且多头注意力机制使得模型在训练过程中关注模型感兴趣的部分, 对于模型的训练有益. 基于以上描述, 本文以建立结构化电子病历为目的, 从电子病历的实体识别开始研究, 提出 WT-MHA-BiLSTM-CRF 模型, 该模型同时考虑文字和标签中的信息. 为保证模型的可信性, 将实验结果与 BiLSTM-CRF[9], BiGRU-CRF, MHA-BiLSTM-CRF 模型进行了对比.

2 电子病历数据预处理

2.1 类别定义

人工将实体分类分为以下 6 个方面: 病症, 身体部位, 手术, 药物, 化验检验, 影像检验. 在测试中发现, 如果直接以中文作为类别标签, 会导致训练结果较差, 因此本文将上述 6 个类别使用英文简写代替, 病症的标签为 DISEASE, 身体部位的标签为 BODY, 手术的标签为 OPERATION, 药物的标签为 MEDICINE, 化验检验的标签为 LAB-CHECK, 影像检验的标签为 IMG-CHECK. 实验期望得到的实体识别效果如图 1 所示.

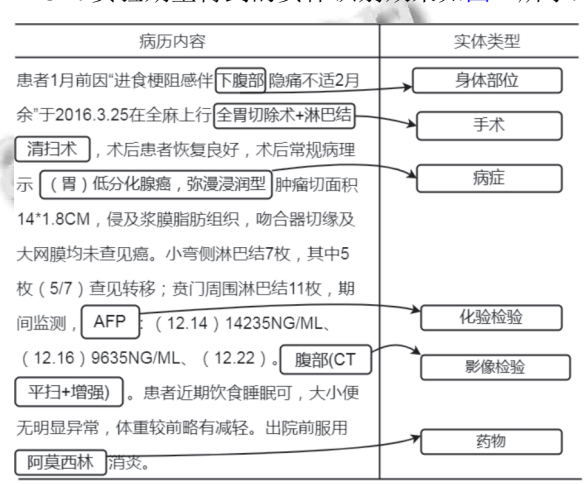

图 1 电子病历实体对应关系图

2.2 原始数据清洗

数据清洗的过程主要利用 Python 程序将原始文本中的空格, 首尾标点符号以及一些错误符号进行删除, 并检查人工标记的起始位置和终点位置的实体是否标记有误, 例如起始位置开始过早或者终点位置结

束过晚等情况. 为了防止标点符号的使用不一致情况, 本文将原始文本中的标点全部使用英文标点表示. 实验发现, 没有进行过数据清洗的数据会比进行过数据清洗的数据在训练结果上产生的误差在 10% 左右, 说明对于原始数据进行清洗在模型训练中存在一定的作用.

3 模型结构

本文提出的一种融合文字与标签的 WT-MHA-BiLSTM-CRF 模型结构如图 2 所示. 在模型中, 输入层获取电子病历的每个文字及文字对应的标签, 将文字和标签分别建立字典, 得到每个字和标签对应在字典中的序号, 即向量化表示后的结果, 将两者进行结合, 输入到 multi-head attention[10] 中, attention 机制会使得模型在句子中捕捉到更多的上下文信息, 其输出结果是在 3 个维度上的加权整合, 并映射到模型设定的矩阵维度上. 该输出的结果作为 BiLSTM 模型的输入, BiLSTM 模型根据上下文信息, 计算每个字符在每一个类别中的概率值大小. 基于 BiLSTM-CRF 模型抛弃了 Softmax 层并使用 CRF 进行代替, CRF 层在训练过程中可以自主学习到一些约束规则, 这些约束可以保证预测标签的合理性, CRF 层将 BiLSTM 层的输出结果作为输入, 进一步约束, 得到更加准确的输出结果, 整体模型将 CRF 的输出作为最后的预测结果, 并与标准结果进行比对, 计算预测的误差.

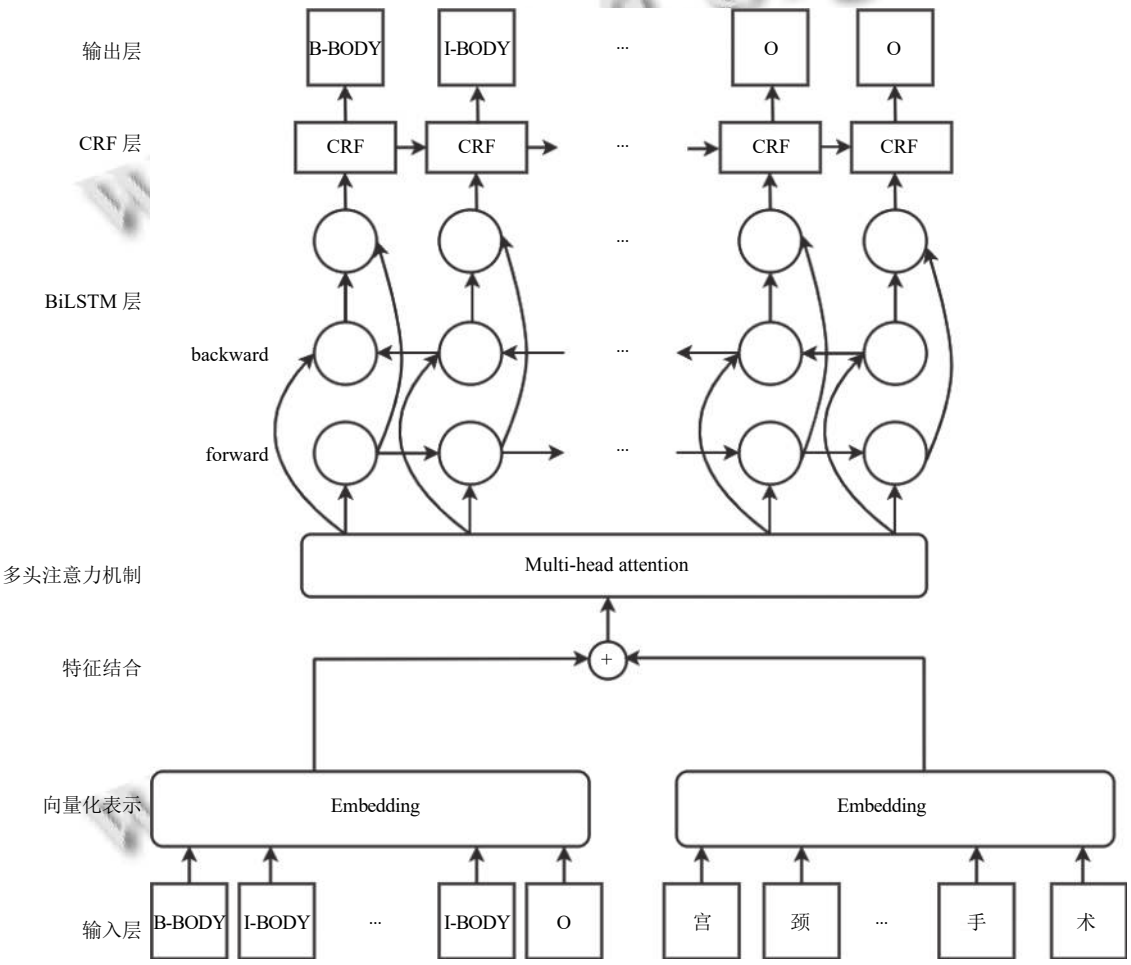

图 2 WT-MHA-BiLSTM-CRF 模型结构图

本文提出的模型相对于传统的命名实体识别模型来说, 同时关注文字和标签中的信息, 使得输入信息更加充分, 发现标签中的隐含信息对模型结果带来的价值.

3.1 BiLSTM 模型

双向长短期神经网络 (BiLSTM)[11] 是一种为解决循环神经网络 (RNN) 中梯度消失或者梯度爆炸问题而提出的模型, 它较好地解决了 RNN 在长依赖训练过程中的缺陷. 长短期神经网络 (LSTM) 是一种单向网络, 主要使用门机制[12] 来解决 RNN 中的长期依赖问题, 其模型结构如图 3 (图中 sig 代表 Sigmoid 函数) 所示.

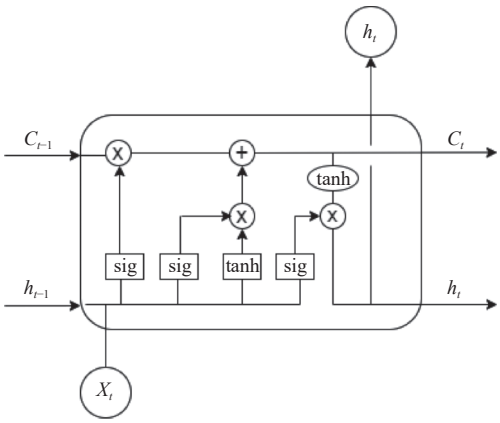

图 3 LSTM 模型结构图

LSTM[13] 在隐藏单元中提出了 3 种门的概念: 输入门, 输出门, 遗忘门. 式 (1) 和式 (2) 体现了输入门与单元状态更新的计算公式:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x] + b_i) \tag{1}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x] + b_C) \tag{2}$$

输入门用来确定当前输入中被保存状态的个数, 其中, i_t为输入门, W、b分别表示权重矩阵和偏置向量, $\sigma(\cdot)$代表 Sigmoid 激活函数, tanh(·) 代表 tanh 激活函数, 其表达式分别为式 (3) 和式 (4) 所示.

$$\sigma(x) = \frac{1}{1+\mathrm{e}^{-x}} \tag{3}$$

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}} \tag{4}$$

输入门通过控制$\tilde{C}_t$来更新状态C_t, 其计算如式 (5) 所示, 其中, f_t代表遗忘门, 将在下面进行说明.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

输出门将会决定传输多少个单元状态作为 LSTM 的当前输出值, 通过计算隐藏节点h_t来计算预测值和下一个时刻的输入, 式中o_t代表输出值.

$$o_t = \sigma(W_o [h_{t-1}, x] + b_o) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

遗忘门负责上一个时刻的单元状态保存多少到当前时刻, 表现为C_{t-1}中哪些特征将被用于计算C_t, 遗忘门f_t中的每个值位于[0,1]中, 计算公式见式 (8).

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x] + b_f\right) \tag{8}$$

单元状态作为 LSTM 中的核心部分, 始终贯穿整个计算, 其计算表达式为:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{9}$$

本文使用的 BiLSTM 模型相对于 LSTM 模型来说, 能够在未来上下文的预测中有较好的表现, 使得训练不仅能够接受前文的序列也可以得到后续序列, 在进行命名实体识别的过程中会有更好的效果.

3.2 GRU 模型

门控循环单元 (GRU) 也是为了解决循环神经网络存在的长期依赖问题而提出的, 模型中使用了更新门和重置门两种门机制, 使得模型在训练效果不变的同时, 计算更加简单, 其计算过程如式 (10)–式 (13) 所示.

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{10}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{11}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \tag{12}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{13}$$

其中, $\sigma(\cdot)$、tanh(·) 同上述描述, 代表哈达马乘积, 用于计算两个阶相同矩阵的对应位置, 并将它们相乘得到一个新矩阵, z_t代表更新门, 其作用是决定多少信息从之前状态h_{t-1}保存到当前状态h_t, 以及在候选状态$\tilde{h}_t$中得到多少信息. r_t代表重置门, 决定了候选状态$\tilde{h}_t$的计算是否依赖于之前状态h_{t-1}.

3.3 Multi-head attention 原理

由于单层的 attention 所包含的信息可能不够支持众多的下游任务, 因此 2017 年谷歌推出的 Transformer 将其堆叠成 multi-head attention (MHA), 其本质是多次的 self-attention[14] 计算. Attention 机制[15] 会使网络在训练过程中更多的注意到输入包含的相关信息, 而对无关信息进行简略, 从而提高训练的准确性.

Self-attention 对输入的每个词向量创建 3 个新向量: *Query*、*Key*、*Value*, 这 3 个向量分别是词向量和 Q, K, V 三个矩阵乘积得到的, Q, K, V 三个矩阵是一个需要学习的参数, 其定义如下所示:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{14}$$

其中, $\sqrt{d_k}$是一个平滑项, 为了保证训练过程中梯度值的稳定. MHA 根据 self-attention 的计算原理, 将原来的一层提升为 h 层 (h 是一个超参数, 在模型中取 8), 其计算过程如图 4 所示.

MHA 定义如式 (15) 所示:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

$$MultiHead(Q,K,V) = Concat(head_1, \cdots, head_h) W^o \tag{15}$$

其中, W_i^Q, W_i^K, W_i^V 是需要训练的参数权重, W^o 属于 $R^{d_{\text{model}} \times hd_V}$, d_{model}代表了模型中 Q, K, V 的维度, $Concat()$ 代表拼接函数, 用于将多层的 Q, K, V 函数拼接起来.

3.4 CRF 模型

条件随机场 (CRF)[16] 是一种较为经典的条件概率分布模型, 通过观测序列$X = (x_1, x_2, \cdots, x_n)$来计算状态序列$Y = (y_1, y_2, \cdots, y_n)$的条件概率值$p(y|x)$, CRF 的简化定义公式如下:

$$P_w(y|x) = \frac{\exp(w \cdot F(y,x))}{Z_w(x)}$$

$$Z_w(x) = \sum_y \exp(w \cdot F(y,x)) \tag{16}$$

其中, w代表权值向量, $F(y,x)$表示全局特征函数. CRF 模型可以自主的学习到文本中的约束, 例如在命名实体识别[17] 中, I-标签的后面不会出现 B-标签, O 标签的后面不会出现 I-标签等, 这些约束会帮助训练结果更加合理.

CRF 模型通过式 (17) 得到评估分数, 式中 $Emit$ 代表 BiLSTM 输出的概率, $Trans$ 代表对应的转移概率.

$$score(x,y) = \sum_i Emit(x_i, y_i) + \sum_i Trans(y_{i-1}, y_i) \tag{17}$$

最后模型使用最大似然法来进行训练, 相应的损失函数为:

$$-\log p(y|x) = -score(x,y) + \log Z(x) \tag{18}$$

由于$Z(x)$无法直接计算, 因此使用前向算法进行推导, 在深度学习框架中可以对损失函数进行求导或者梯度下降的方法来优化, 使用 Viterbi 算法进行解码, 从而找到最优结果.

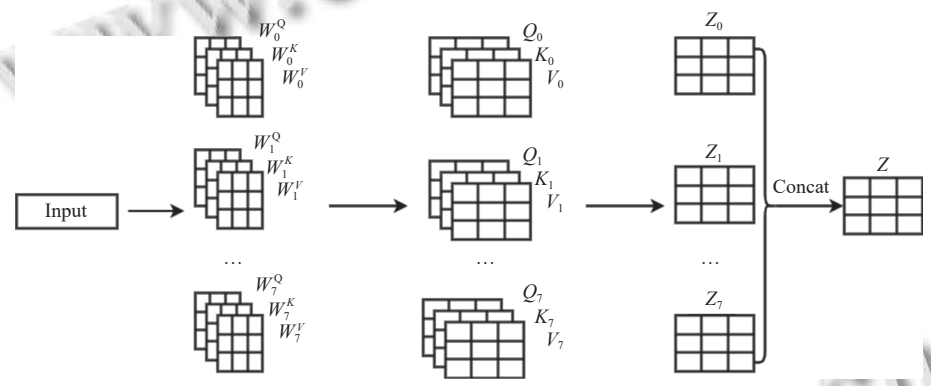

图 4 Multi-head attention 计算过程图

4 实验及结果分析

4.1 实验数据集

实验过程中比较了 4 种模型的训练情况, 使用人工标注过的 1 000 份数据集, 按照 6:2:2 的比例将数据划分为训练集、验证集、测试集. 在实体识别中常用的实体标注方式[18] 包括 BIO、BIOE、BIOES 等, 本实验使用 BIO 标注方式, 具体的标注情况与实体数量见表 1 和图 5 所示. 根据第 2.2 节中所述, 数据集的准确性对模型最终结果存在影响, 因此实验对原始数据进行预处理, 主要包括中英文统一, 利用程序检查原始数据集中是否有启始位置错误信息, 并针对错误信息进行修改.

4.2 数据集评价指标

实验使用精确率 (结果中的 P 值)、召回率 (结果中的 R 值) 和 $F1$ 值作为模型评价指标, 其中精确率代表在所有预测结果中为正确的个数在实际正确分类所占的比例; 召回率代表所有预测正确的结果在实际正确中的占比; $F1$ 值使用加权和平均来保证两者在结果中的作用, 由于综合考虑了精确率和召回率的结果, 因此 $F1$ 的值往往作为实验结果的最有力的证明, 其值越高说明实验的效果越好, 三者的计算过程见式 (19)–式 (21).

$$P = \frac{T_p}{T_p + F_p} \times 100\% \tag{19}$$

$$R = \frac{T_p}{T_p + F_N} \times 100\% \tag{20}$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{21}$$

其中, T_p表示在测试集中正例被正确分类的个数, F_p代表负例被错误分类为正例的个数, F_N表示正例被错误分类为负例的个数.

表 1 数据标注表

类别	开始标记	中间标记	实体数量
病症	B_DISEASE	I_DISEASE	4212
身体部位	B_BODY	I_BODY	8426
手术	B_OPERATION	I_OPERATION	1492
药物	B_MEDICINE	I_MEDICINE	1931
化验检验	B_LAB-CHECK	I_LAB-CHECK	1195
影像检验	B_IMG-CHECK	I_IMG-CHECK	972

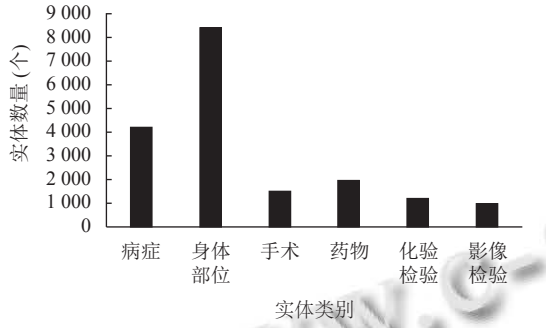

图 5 实体数量图

4.3 实验环境与结果

本文中命名实体识别模型基于 PyTorch 框架, 详细实验环境设置见表 2 所示.

表 2 训练环境表

项目	环境
操作系统	Windows 7
CPU	i5-5200U@ 2.20 GHz
GPU	Nvidia Tesla K80
Python版本	3.7.12
PyTorch版本	1.10.0+cu111

训练中模型的详细参数为: 每次训练选取的样本数 batch_size 值为 64, 学习率 lr 设定为 1E–4, 训练轮数 epoch 设置为 30, 时序模型的网络层数设置为 1, 词向量维数为 128, 使用 Adam 优化器, 丢失率 dropout 值为 0.5.

本文对 4 种模型进行了测试来增强模型结果的说服力, 4 种模型定义如下.

(1) BiLSTM-CRF: 使用单层的 BiLSTM 网络, 并将模型的输出结果直接作为条件随机场的输入计算最终结果.

(2) BiGRU-CRF: 使用单层的 BiGRU 模型, 将输出结果投入 CRF 计算.

(3) MHA-BiLSTM-CRF: 对字向量做 Embedding 后先进入 mutli-head attention 模型中, 模型输出的结果作为 BiLSTM 模型的输入, 其结果作为 CRF 层输入.

(4) WT-MHA-BiLSTM-CRF: 同时对字、标签做 Embedding, 将两者结合, 作为 mutli-head attention 的输入, 之后进入 BiLSTM 模型训练, 最终进入 CRF 层进行训练.

表 3 展示了 4 种模型在各个实体上的精确率、召回率和 $F1$ 值, 从结果来看, WT-MHA-BiLSTM-CRF 在大多数的实体分类结果上均有一定的提高, 与 BiLSTM-CRF 模型的比较中, 在病症方面 $F1$ 值提高了 3%, 身体部位提高了 6%, 手术提高了 4%, 化验检验提高了 7%, 影像检验提高了 11%. 但是在药物的类别方面, WT-MHA-BiLSTM-CRF 的结果与其他模型相比有些下降, 初步分析原因是, 由于本文提出的模型是针对字和标签同时作为模型的输入来进行训练, 并且经过 attention 机制的处理, 在实验过程中, 可能过度的对某些字、标签产生了过度的关注, 而使得分类结果产生了误差, 因而导致分类效果的下降, 在以后的研究中会针对这一方面进行深入研究. 最后, 图 6 展示了 WT-MHA-BiLSTM-CRF 模型在各个实体中的预测结果.

表 3 实验结果表 (%)

模型	BiLSTM-CRF			BiGRU-CRF			MHA-BiLSTM-CRF			WT-MHA-BiLSTM-CRF		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
病症	80.41	83.13	81.75	80.08	86.52	83.14	79.36	87.71	83.31	85.25	84.71	84.87
身体部位	87.95	72.39	79.41	89.67	74.68	81.42	82.47	80.94	81.70	83.48	87.48	85.50
手术	87.12	83.96	85.49	87.19	86.32	86.75	90.45	87.46	88.93	90.28	88.29	89.27
药物	95.44	88.05	91.60	92.14	93.76	92.94	96.46	92.92	94.67	94.58	87.44	90.86
化验检验	81.39	77.25	79.27	84.30	82.95	83.62	78.82	80.56	79.67	88.96	84.50	86.67
影像检验	78.79	78.67	78.73	89.49	87.00	88.19	83.99	85.91	84.86	84.75	95.79	89.86

5 结论与展望

本文提出了一种结合字与标签同时训练的 WT-MHA-BiLSTM-CRF 模型, 用于解决结构化电子病历中命名实体识别的问题. 传统的 BiLSTM-CRF 模型更多的注重字在模型中训练产生的结果, 忽略了标签中可能隐含的信息, 因此在训练输入之前, 将字和标签同时做 Embedding 处理. 并使用 mutli-head attention 使模型更多的注意力放在关键的字和标签上, BiLSTM 结合

上下文得到每个字的类别概率, 模型的最后使用 CRF 对分类结果进行处理, 使得模型的预测输出更有说服力. 从实验结果中可以看出, 该模型在大部分的类别识别中均有提升. 对于个别类别来说, 考虑到模型的结果中可能有部分信息被过度关注, 导致训练结果上有些许问题, 在以后的研究中也会针对这个问题着重来处理.

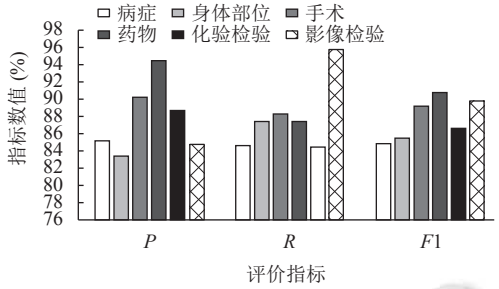

图 6 WT-MHA-BiLSTM-CRF 模型各类别结果图

参考文献

1 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. 软件学报, 2016, 27(11): 2725–2746. [doi: 10.13328/j.cnki.jos.004880]

2 孟捷. 基于中文电子病历文本的医学语义网络构建方法研究 [硕士学位论文]. 北京: 北京交通大学, 2019.

3 江涛. 基于深度神经网络的电子病历命名实体识别关键技术研究与应用 [硕士学位论文]. 成都: 电子科技大学, 2020.

4 沈宙锋, 苏前敏, 郭晶磊. 基于 XLNet-BiLSTM 的中文电子病历命名实体识别方法. 智能计算机与应用, 2021, 11(8): 97–102. [doi: 10.3969/j.issn.2095-2163.2021.08.021]

5 王若佳, 赵常煜, 王继民. 中文电子病历的分词及实体识别研究. 图书情报工作, 2019, 63(2): 34–42.

6 Zhang D, Chi CY, Zhan XG. Leveraging lexical features for Chinese named entity recognition via static and dynamic weighting. IAENG International Journal of Computer Science, 2021, 48(1): IJCS_48_1_13.

7 Ji B, Li SS, Yu J, *et al*. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models. Journal of Biomedical Informatics, 2020, 104: 103395. [doi: 10.1016/j.jbi.2020.103395]

8 李丹, 徐童, 郑毅, 等. 部首感知的中文医疗命名实体识别. 中文信息学报, 2020, 34(12): 54–64. [doi: 10.3969/j.issn.1003-0077.2020.12.009]

9 张旭, 朱艳辉, 梁文桐, 等. 基于 SoftLexicon 的医疗实体识别模型. 湖南工业大学学报, 2021, 35(5): 77–84. [doi: 10.3969/j.issn.1673-9833.2021.05.010]

10 吕晴, 赵奎, 曹吉龙, 等. 基于文本与图像的肺疾病研究与预测. 自动化学报, 2022, 48(2): 531–538.

11 屈倩倩, 阚红星. 基于 Bert-BiLSTM-CRF 的中医文本命名实体识别. 电子设计工程, 2021, 29(19): 40–43, 48.

12 Ma RT, Peng ML, Zhang Q, *et al*. Simplify the usage of lexicon in Chinese NER. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020. 5951–5960.

13 Gong C, Li ZH, Xia QR, *et al*. Hierarchical LSTM with char-subword-word tree-structure representation for Chinese named entity recognition. Science China Information Sciences, 2020, 63(10): 202102. [doi: 10.1007/s11432-020-2982-y]

14 Yin JZ, Luo SL, Wu ZT, *et al*. Chinese named entity recognition with character-level BLSTM and soft attention model. Journal of Beijing Institute of Technology, 2020, 29(1): 60–71.

15 曾青霞, 熊旺平, 杜建强, 等. 结合自注意力的 BiLSTM-CRF 的电子病历命名实体识别. 计算机应用与软件, 2021, 38(3): 159–162, 242. [doi: 10.3969/j.issn.1000-386x.2021.03.024]

16 丁锋, 孙晓. 基于注意力机制和 BiLSTM-CRF 的消极情绪意见目标抽取. 计算机科学, 2022, 49(2): 223–230.

17 肖丹. 基于半监督学习的中文电子病历实体关系抽取研究 [硕士学位论文]. 绵阳: 西南科技大学, 2021.

18 袁贞明, 沈辉, 俞凯, 等. 基于电子病历文本的诊疗事件实体抽取研究. 中国数字医学, 2021, 16(7): 33–38. [doi: 10.3969/j.issn.1673-7571.2021.07.007]

(校对责编: 孙君艳)