文章编号：1003-0077(2020)12-0054-11

部首感知的中文医疗命名实体识别

李 丹[1,2]，徐 童[1,2]，郑 毅[3]，王喆锋[3]，陈恩红[1,2]

(1. 大数据分析与应用安徽省重点实验室(中国科学技术大学)，安徽 合肥 230027；
2. 中国科学技术大学 计算机科学与技术学院，安徽 合肥 230027；
3. 华为技术有限公司，浙江 杭州 310052)

摘 要：人工智能技术的发展推动了医疗领域的智能化，为提升医疗效率、改善医疗水平提供了新的助力。同时，这一新的趋势也催生了海量的电子病历文本，其所蕴含的丰富信息具有巨大的潜在挖掘与应用价值。然而，当前中文电子病历的命名实体识别研究工作并没有全面考虑中文及中文医疗领域的特殊性，而是将面向通用数据集的模型迁移到医疗领域的实体类型中，分析效果较为有限。针对这一问题，该文设计了长短期记忆网络与条件随机场的联合模型并引入 BERT 模型；在此基础之上，考虑到医疗领域命名实体鲜明的部首特征，通过将部首信息编码到字向量中，并且结合部首信息修改条件随机场层得分函数的计算方式，有效地提升了医疗领域命名实体的抽取能力。通过两项电子病历数据集的实验结果表明，该文提出的模型整体效果略高于通用的实体识别模型，并对疾病诊断等特定类型的实体词的识别效果具有较为明显的提升。

关键词：命名实体识别；长短期记忆网络；条件随机场；BERT

中图分类号：TP391 **文献标识码**：A

Radical-Aware Named Entity Recognition for Chinese Medical Records

LI Dan[1,2], XU Tong[1,2], ZHENG Yi[3], WANG Zhefeng[3], CHEN Enhong[1,2]

(1. Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China, Hefei, Anhui 230027, China;
2. School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China;
3. Huawei Technologies Co.Ltd, Hangzhou, Zhejiang 310052, China)

Abstract: The general named entity recognition fails to capture the features in Chinese characters as well as Chinese medical records. In this paper, we integrate the BERT into a joint model of bi-directional long short-term memory and conditional random fields for better performance. Considering the unique feature of radicals for medical entities, we encode the radical information into the word vector, and then modify the scoring function of the CRF layer. Experiments on two real-world electronic medical record datasets validate that the proposed method outperforms the state-of-the-art baseline methods, especially for the disease-related named entities.

Keywords: named entity recognition; long short-term memory; conditional random field; BERT

0 引言

随着互联网技术的飞速发展，网络信息呈现指数级增长的态势，一大批在线医疗社区和医疗信息问答网站也随之涌现，使得海量的医疗诊断信息以电子文档的形式呈现在人们面前。据统计，仅国内的某寻医问药网站就包含了 2004 年 11 月至今十余年的疾病问答诊断数据，从而形成了海量且具有巨大潜在价值的医疗数据。然而，与数据库不同的是，

收稿日期：2019-10-09 **定稿日期**：2019-12-04

基金项目：国家重点研发计划(2018YFB1004300)；国家自然科学基金(U1605251，61703386)；中央高校基本科研业务费专项资金(WK9110000014)；安徽省重点研发计划项目(1804b06020377)

这些医疗数据文本大多处于非结构化的状态。为了充分利用这些医疗领域文本蕴含的信息，通过命名实体识别技术有效抽取其中有用的医疗实体词，已成为实现智慧医疗的前提和基础。

目前，主流的命名实体识别技术主要包括基于双向长短期记忆网络（以下简称 BiLSTM）与条件随机场（以下简称 CRF）的识别模型[1]，以及基于 BERT 的预训练—微调两阶段模型[2]等。在此之前，还有基于规则的识别方法[3]、基于 CRF 的识别方法[4]、基于隐马尔科夫模型（以下简称 HMM）的识别方法[5]，以及基于 LSTM 的识别方法[6]等传统技术。在这些模型之中，整体识别效果最好的为采用深度学习技术的 BiLSTM＋CRF 模型和 BERT 模型。这两者能够基于神经网络提取出包含上下文语义的字向量特征，进而对每个字符给出较为合理准确的标签预测。然而，这些模型也存在有待改进的地方。LSTM 模型的通用性，以及 BERT 发布的预训练模型采用的中文数据集的通用性，使得其比较适合通用的实体识别数据集上的任务，提取的特征也仅仅局限于上下文信息与标注的语法正确性，并没有根据医疗领域电子病历数据集的实体词特征来调整模型以适应该类数据集，从而使得传统模型在该类数据集上的识别效果并不理想。

更为重要的是，医疗领域电子病历的实体词之间往往存在着一些包含关系，因此很难采用传统的分词技术先分词再进行基于词的实体识别。例如，“左心室瓣膜损伤”一词，正确标注为“疾病信息”，而如果采用中文分词技术，则会被分词为“左心室”“瓣膜”“损伤”三个实体词，这三个实体词将被分别标注为“身体部位”“身体部位”“症状”，从而得到了与正确标注相悖的结果。

与此同时，我们注意到在中文电子病历数据集中，一些特定类型的实体词往往具有不同于通用实体词的特征，尤其体现在特定实体所具有的部首种类上。例如，许多组成疾病实体词的字往往具有“疒”部首，例如，“疤”“痫”“痛”等，而许多组成身体部位的实体词的字往往具有“月”部首，例如，“肌”“骨”“肾”等，这些部首信息在确定标注时也具有非常重要的参考价值，特别是在由多个组件所构成的医学实体，如“身体部位＋症状”格式的疾病信息等方面将起到促进作用。然而目前，这一信息尚未被通用命名实体识别模型充分利用。

针对上述问题，本文将从通用命名实体识别模型出发，同时把部首信息作为一个重要的特征加入到现有的模型中。通过在字向量中增加部首编码，以及在 CRF 层计算标注得分时将部首信息考虑在内，形成了面向中文电子病历的特定命名实体识别模型。基于两个数据集的实验表明，本文所提出的方法相比传统的通用模型有着更好的识别效果，特别是对疾病诊断等特定类型的实体词的识别效果具有较为明显的提升。

本文组织结构如下：第 1 节对中文电子病历上的命名实体识别问题进行定义，并简要介绍传统的通用模型，第 2 节介绍了在此基础上对传统的通用模型所做的修改，第 3 节介绍实验方案与实验结果。最后对全文进行总结并展望未来的工作。

1　相关工作

在本节中，我们将简要介绍与中文医疗命名实体识别相关的研究进展，并着重介绍两个当前主流的通用命名实体识别模型，即 BiLSTM＋CRF 与 BERT 模型的结构及其原理。

1.1　非深度学习模型介绍

常见的非深度学习模型以隐马尔可夫模型（HMM）和条件随机场（CRF）为主。

在 HMM 模型中，将字符的标注定义为隐状态，将字符的值定义为观察序列。HMM 假定一个句子由以下方式生成：①每一对隐状态 i，j 之间以转移概率 a_{ij} 进行转移；②每一个隐状态 i 以发射概率 b_{ik} 产生观察序列中的状态 k。一个观察序列对应的一个隐状态序列概率为各发射概率和转移概率的累乘，其反映了该观察序列由此隐状态序列产生的可能性大小。在建模时，转移概率和发射概率通过对大量样例数据的统计计算得到，在预测时，通过 Viterbi 解码的方式得到使得马尔可夫链的概率和发射概率累乘最大化的隐状态序列，也即观察序列对应的标注。

在 CRF 模型中，字符的标签属性由字符词性、前缀、上一个字符的词性等特征决定，并认为这些特征在决定标签时，相应的权重也是不同的。定义特征函数 $f(s,i,l_i,l_{i-1})$ 的含义为：对于句子 s，若其第 i 个字的属性满足 l_i 且第 $i-1$ 个字的属性满足 l_{i-1}，则该特征函数取值为 1，否则取值为 0。将一个句子标注为一个标签序列的得分为其各个特征函数的加权求和，得分越高代表该句子属于此标注序列的可能性越大。在建模时，通过使得正确标注的

得分在所有可能的标注序列和中占比最大，即可学习得到各特征函数的权重。在预测时，同样地通过Viterbi解码的方式来获得使得分最大的标注序列。

1.2 BiLSTM+CRF模型介绍

长短期记忆网络(long short term memory，LSTM)是一种特殊的递归神经网络(RNN)，其解决了RNN学习长期依赖困难的问题，在很多序列问题上都取得了巨大的成功，得到了广泛使用。LSTM结构内部存在三个门(gate)[7]，分别为输入门(input gate)，遗忘门(forget gate)和输出门(output gate)。遗忘门用于决定从神经网络细胞状态中丢弃什么信息，输入门决定哪些新信息被存放在细胞状态中，输出层基于过滤后的新细胞状态决定输出的值。如式(1)～式(4)所示。

$$i_t=\delta(\boldsymbol{W}_{xi}x_t+\boldsymbol{W}_{hi}h_{t-1}+\boldsymbol{W}_{ci}c_{t-1}+b_i) \quad (1)$$

$$f_t=\delta(\boldsymbol{W}_{xf}x_t+\boldsymbol{W}_{hf}h_{t-1}+\boldsymbol{W}_{cf}c_{t-1}+b_f) \quad (2)$$

$$c_t=f_tc_{t-1}+i_t\tanh(\boldsymbol{W}_{xc}x_t+\boldsymbol{W}_{hc}h_{t-1}+b_c) \quad (3)$$

$$o_t=\delta(\boldsymbol{W}_{xi}x_t+\boldsymbol{W}_{hi}h_{t-1}+\boldsymbol{W}_{ci}c_{t-1}+b_i) \quad (4)$$

其中，h_t 代表 t 时刻的隐藏层，x_t 代表 t 时刻的输入，c_t 代表 t 时刻的细胞状态，o_t 代表 t 时刻的输出，f_t 表示遗忘门的输出，i_t 代表 t 时刻输入门的输出，$\boldsymbol{W}$ 表示各权重参数矩阵。图1展示了一个基础的LSTM序列标注模型，出于方便，将LSTM的内部结构单元予以简化表示。对于输入语句的每一个字符，LSTM输出其被标注为各个标签的可能性，可能性最大的标签即为输出的预测标签。

在序列标注任务中，当对一个词的标签进行预测时，不仅要结合当前词之前的输入的历史特征，为了提高效果，当前词的后文特征信息也要考虑在内。基于此，可以利用LSTM的特点构建一个BiLSTM网络(bidirectional LSTM networks)[8-10]，前向LSTM的输入为正常顺序的语句，反向LSTM的输入则依次为逆置后的词序列。最终将前向LSTM和反向LSTM的输出拼接起来作为预测的可能性。

但通过上文方式提取出的向量，虽然包含了丰富的特征信息，但仍缺少在语法层面上的关注，即只结合了上下文的特征信息，没有利用历史标注信息来减少模型产生不合乎语法的标注序列的可能性。为了实现这一特性，需要将上层网络接入一个CRF层。CRF层的网络参数是一个转移得分矩阵 T，其中 T_{ij} 表示从一个标签 i 转移到当前标签 j 的转移得分。至此，一句长为 L 的句子 X，将其标注为标

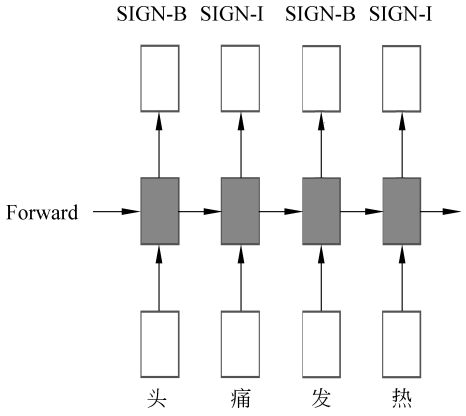


图1 LSTM用于序列标注

签序列 I 的得分函数 score(X，I)的计算如式(5)所示[11]。

$$\begin{aligned}&\text{score}([X]_1^L,[I]_1^L)\\&=f_{1,[I]_1}+\sum_{t=2}^{L}(f_{t,[I]_t}+T_{[I]_{t-1},[I]_t})\end{aligned} \quad (5)$$

其中，f 是下游BiLSTM网络输出的每个字符对每种标签的评分，f_{ij} 代表第 i 个字符被标注为标签 j 的标注得分。最终的得分函数为BiLSTM的评分与转移得分之和。在训练时，通过最大化真实标注序列的得分在所有可能的序列得分中占的比重来更新各参数；预测时根据 f 得分和转移矩阵 T，可以通过Viterbi解码得到最优标注序列。最终，基于BiLSTM+CRF的模型如图2所示[12-14]。

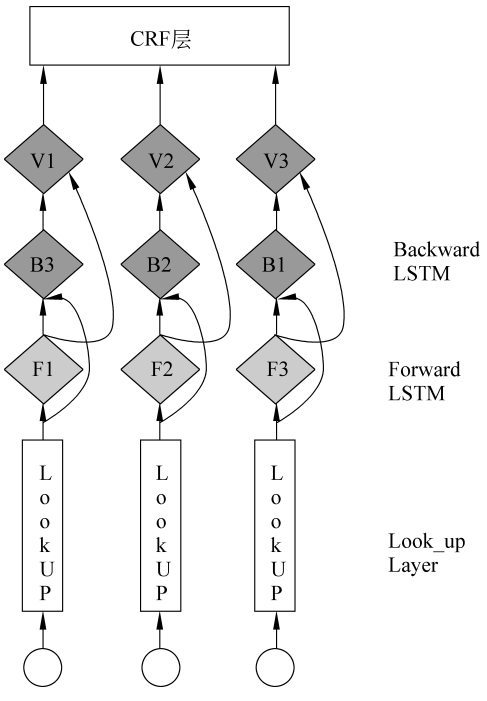


图2 基于BiLSTM+CRF的命名实体识别模型

1.3　基于 BERT 的命名实体识别简述

BERT 于 2018 年被提出，其采用基于 transformer 的双向编码器表征，能够高效地结合上下文抽取特征信息，并且产生类 ELMO 的词向量；同时基于 pre-training 与 fine-tuning 的二阶段模型可以利用预训练好的参数减少训练代价。由于 Google 团队已经开源了 BERT 的代码，并且提供了预训练的模型，因此本文的模型是在加载其基于中文语料库训练好的模型 Chinese_L-12_H-768_A-12 的基础上，通过参数微调来实现 BERT 提取特征的。

传统基于 BERT 的命名实体识别模型流程如图 3 所示。输入的字符被转换为 Token，并为每个句子加上开始及结束标志，模型的输出即为每个 Token 的最有可能的预测标签。

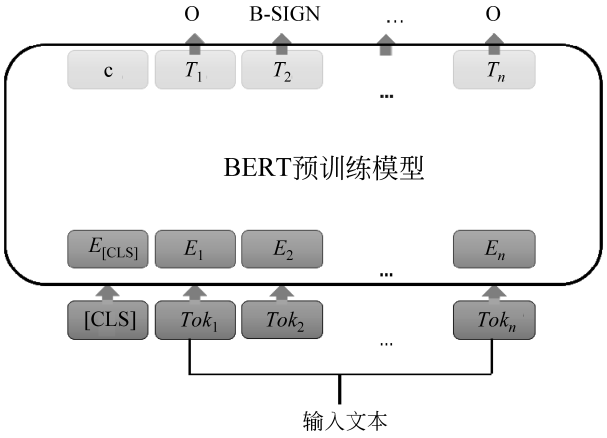


图 3　基于 BERT 的命名实体识别

2　引入部首特征的医疗命名实体识别模型

引言中提到，传统命名实体识别模型是针对通用数据集的，底层网络提取的特征局限于上下文的语义层面特征，而缺少对特定领域的数据集，例如对中文电子病历的特有特征的关注。就像英文中可以根据单词的词根、词缀来猜测意义和性质一样，中文的部首和笔画也蕴含着大量字义信息，而灵活运用这些信息可以对模型的效果有所提升。

本节将着重介绍如何表征中文部首特征，尤其是与疾病和症状相关的部首，并将其与本文所提出的医疗命名实体识别模型相结合，从而有效提升实体识别效果。

2.1　中文电子病历数据集的部首特征

中文电子病历数据不同于通用数据集的地方在于其领域的特殊性，从而使得该类数据中组成实体词的字符也大多是限定且具有特殊特征的。如引言中所述，一些疾病实体词的汉字大都有“疒”部首，组成身体部位实体词的汉字大都具有“月”部首。当然，组成医疗领域实体词的汉字部首特征并不局限于这两种。例如，中文的五行“金木水火土”也常常被用来作为医学领域各实体词的特征部首，如金对应着身体中的常见微量元素及一些药品名，例如，“×××钠”“×××钙”等；木对应的“×××术”“查体×××”“×××栓”“椎”，以及一些中成药物的药品名等；水对应着各种体液（血浆、组织液、淋巴液）及代表一些症状的“湿”“滑”“溢”“溶”“溃”等；火对应的各种炎症的结尾字符，以及一些症状实体字符，如“烫”“灶”“烂”；土对应的各种症状，表示“垂”“均”“型”“塞”以及身体部位修饰词“壁”“切”等。此外，还有其他的一些诸如“心”“阝”“车”“禾”“扌”“气”“口”等部首，类似的这些部首对于鉴定医疗实体词也具有重要的参考价值[15]。

然而，实际使用这些部首的过程面临以下两个问题：①存在着一些冗余部首。例如，“阝”与“耳”部首表达的都是同一种释义，却是两个不同的部首，同样的还有“忄”与“心”“火”与“灬”等。这样的冗余存在既增大了训练代价，也降低了部首信息的表达能力。②一些释义截然不同的字符有着相同的部首。一个很直观的样例是“朝”与“脚”，这两个字的部首都是“月”，然而前者的释义为“早晨”或“朝代”，后者的释义为“身体部位”；这种多个释义字对应一个部首的情况也会对部首的特征表达能力产生影响。

通过查阅相关汉语汉字文献[16-18]可以发现，简体中文汉字往往都是由繁体中文汉字简化演变而来，繁体汉字相比前者在字形字构与部首上，往往更具有解释性。并且，采用繁体部首可以很好地解决上述提到的两个问题。以“阝”为部首和“耳”为部首的汉字的繁体部首都是“耳”，以“忄”与“心”为部首的汉字的繁体部首都是“心”，以“火”与“灬”为部首的汉字的繁体部首都是“火”。这在很大程度上解决了冗余部首的问题。同时，继续沿用上例中的“朝”和“脚”，“朝”的繁体部首是“月”，代表“月亮”；而“脚”的繁体部首是“肉”，象征着身体。这就与该汉字本来的释义非常接近，也缓解了多个释义字对应一个部首引起的部首表达能力降低的问题。因此，如果能为数据集中的汉字找到相应的繁体部首，就能较好地提取出部首层面的特征，并以部首特征作为一个汉字重要的释义表达。

本文的繁体部首是通过爬取在线新华字典页面①所得到的。在线新华字典收录了 7 万多个汉字的所有信息，包括部首、繁体部首(如果存在的话)等信息。本文将电子病历数据集中的所有汉字依次经由在线新华字典进行查询，构造了一个汉字-部首，或者是汉字-繁体部首(如果存在繁体部首)的键—值对字典。从数据预处理的情况来看，除了特殊的非汉字字符与标点以外，数据集中所有的汉字都能在新华字典中找到。

2.2 部首信息的编码与向量嵌入

为实现将部首信息应用到中文医疗实体抽取任务，首先需要将部首信息融入到下游网络的特征抽取中。从实践效果来看，BiLSTM 和 BERT 都是比较优秀的提取特征的模型，并且 BERT 抽取出的类 ELMO 向量[19]往往有着更好的表达能力。因此，本文将基于这两种技术所生成的字向量来实现部首信息的嵌入。

整体上看，下游网络的字向量特征由两部分组成：①每个字符经查找层，与该字符的部首经查找层得到的向量拼接，再经过 BiLSTM 网络得到的字向量。②BERT 模型得到的字向量表示。最终，将这两部分向量拼接起来作为最终提取的字向量，抽取字向量的模型结构如图 4 所示。查找层矩阵的行向量不仅包括数据集中的字符，还增加了所有字符的部首。每个字符首先通过前文构建的汉字-部首映射字典得到其对应的部首，然后将字符与对应的部首分别经过同一个可训练的查找层矩阵，将得到的两部分向量拼接起来，记为 x_a，作为经过查找层的包含了部首信息的初步表示。将 x_a 作为 BiLSTM 网络的输入，得到的输出记为 x_b。对于 BERT 而言，由于它是端到端的模型，而本文只需要它提取的词向量，记为 x_c。将 x_b 与 x_c 拼接起来，即是最终整个下游网络提取的字向量表示。

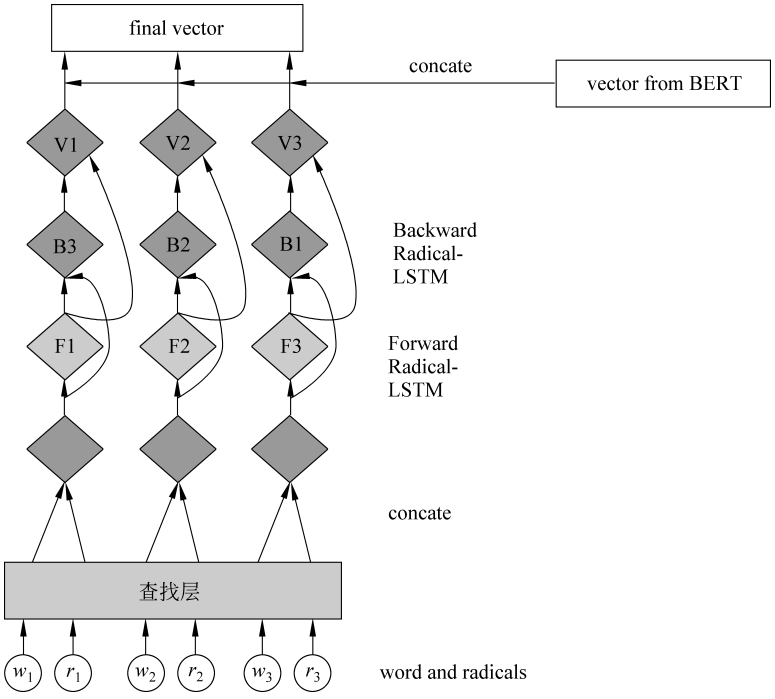


图 4 提取字向量的模型结构

① https://zd.hwxnet.com/

需要引起注意的是，类比于通过抽取各个字与其上下文的信息，也就是传统命名实体识别抽取的向量来作为字符信息的表示，部首层面的上下文部首信息同样也是非常重要的。因此，如前所述，本文的做法是将查找层得到的字符向量与对应的部首向量拼接起来，经过 BiLSTM，而不是仅仅字符向量经过 BiLSTM 再与部首向量拼接。这样做的好处是还可以学习到一些常见的医疗实体词的组成字的部首组合规律，从而更为充分地利用部首信息特征。

2.3　结合部首信息修改 CRF 层

除了将部首信息融入到下游网络的特征抽取之外，本文还针对特征对 CRF 层做了相应的修改。原始的 CRF 层如前所述，得分由下游网络输出的每个字被标注为每个标签的得分 f 与转移矩阵带来的转移得分两部分组成。考虑转移矩阵的原因是为了综合利用历史的标注信息，以减少出现不合法的标注序列（例如，I 类标签在 B 类标签之前）的情况。还需要注意的是，通用模型的数据标注类型既可以为 B-I-O 标注，即一个类别的实体词的开始字符加 B-前缀，其他字符加 I-前缀，非实体词为 O 标注。同时，也可以为 B-I-E-S-O 标注，即一个类别的字符数大于 1 的实体词的开始字符加 B-前缀，其他非结束字符加 I-前缀，结束字符加 E 前缀，字符数为 1 的实体词加 S-前缀，非实体词为 O 标注。在通用模型上，两种标注表示法的效果差异通常并不明显。

借鉴 CRF 的转移矩阵在加强标注的语法正确性起到的作用，本文在 CRF 层增加了一个部首—标签矩阵 R，用来加入部首在确定字符的标注时的影响。R_{ij} 代表将部首索引为 i 的字符标注为标签索引 j 的得分，从而，新的加入了部首特征的 CRF 的标注得分的计算方式应该再加入新的部首—标签得分。一个长为 L 的句子 X，将其标注为标签序列 I 的得分函数 score(X，I)的更新后的计算如式(6)所示。

$$\begin{aligned}\text{score}([X]_1^L,[I]_1^L)=&f_{1,[I]_1}+R_{r(1),[I]_1}\\&+\sum_{t=2}^{L}(f_{t,[I]_t}+T_{[I]_{t-1},[I]_t}\\&+R_{r(t),[I]_t})\end{aligned}\tag{6}$$

其中，f 是下游网络特征提取经过线性变换层和 softmax 层所输出的每个字符对每种标签的评分，f_{ij} 代表第 i 个字符被标注为标签 j 的标注得分。$r(i)$代表第 i 个字符的部首的索引，T 为转移矩阵，R 为部首标签矩阵。最终的标注得分由三部分组合而成：下游网络输出得分、标签转移得分、部首—标签得分。

对于部首—标签矩阵，初始化的值应均匀设置，表示着初始状态下各个部首的字符被标注为各个标签的得分是一样的，这样可以在即使某些部首对标签不敏感的情况下也不会出现太大的偏差。同理，损失函数的计算与 Viterbi 解码过程中，都只需要将原始的得分函数计算方式修改为新的 CRF 得分函数计算方式，其中 f 和 R 可以事先加和作为 f'，这样只需要将原始 CRF 的各个模块的 f 替换为 f'就可以方便地完成修改与计算。

传统的命名实体识别模型对标注的类型并不敏感。但在 CRF 层添加了部首—标签矩阵之后，结合中文电子病历数据集的特征，直观上看 B-I-E-S-O 标注方式应该有更好的效果。经过对电子病历数据集中各实体词的统计发现，一些部首在实体词中出现的位置分布差异比较大，例如，“火”部首大多数情况下出现在实体词末尾，例如，“×××炎”“发热”“发烧”等，“肉”部首比较倾向于出现在实体词开头或者是以单独的字作为实体词等。因此在引入部首—标签矩阵的情况下，采用 B-I-E-S-O 标注的方法能更好地将这种部首在位置分布上的特征考虑进来，从而提高识别的效果。

然而，由于 CRF 层的转移矩阵 T 具有(标签数×标签数)的小规模，而部首—标签矩阵 T 的规模是(部首数×标签数)，这个规模在部首数上千时非常庞大，使得在动态规划进行预测解码时代价也比较大。因此为了避免增大训练负担，本文只选用了一系列特征比较明显的部首集合，记为 C。本文精心选取了 20 个部首，包括“疒”“肉”“金”“木”“水”“火”“土”“心”“耳”“口”“手”“车”“禾”“草”“人”“糸”“气”“大”“血”“尸”等添加到 C 中，维持 T 矩阵与 R 矩阵在同数量级的规模。具体的实现方式如下：当一个字符序列需要预测其标注，在计算序列的得分函数时，只有部首在集合 C 中的字符才会累加其部首—标签得分，否则忽略其部首—标签得分。通过这样的方式，可以有效地控制加入的矩阵的规模与训练代价，同时保留了特征明显的部首在标注时的影响力。

3　实验结果与分析

在本节中，将介绍基于两个测评数据集的实验

对比情况及在不同类型实体上的效果。

3.1 实验设置

本文的实验数据集采用CCKS2017[①]与CCKS2019[②]有关测评的电子病历数据集，其中2017版本的数据集包含5种类别共计4 304个实体：疾病和诊断(DISEASE)、治疗(TREATMENT)、症状和体征(SIGN)、检查和检验(CHECK)、身体部位(BODY)。该语料集涵盖了病史特点、出院情况、一般项目、诊疗经过四个项目，其中每个项目各包含300条病历项目，折合30多万字。数据集的各实体分布如表1所示。CCKS2019版本的数据集包含6种类别共计1 156个实体：手术(OPERATION)、身体部位(BODY)、药物(MEDICINE)、影像检查(IMAGE_CHECK)、疾病和诊断(DISEASE)、实验室检查(LAB_CHECK)。CCKS2019数据集的各实体分布如表2所示。相比2017版本的数据集，该数据集规模更小，类别更多，所以预期的标注效果也会有所下降。

实验采用P(准确率)、R(召回率)、F值来衡量模型识别效果的优劣。具体的衡量手段为：统计测试集的每个句子中人工标注的实体词数A_1、模型识别出来的实体词数A_2、识别正确的实体词数TP，那么$P=\mathrm{TP}/A_2$，$R=\mathrm{TP}/A_1$，F为两者的调和平均。

表1 CCKS2017数据集各实体分布

实体类别	个数	占比
TREATMENT	180	0.04
DISEASE	123	0.02
CHECK	1417	0.33
SIGN	1098	0.26
BODY	1486	0.35

表2 CCKS2019数据集各实体分布

实体类别	个数	占比
OPERATION	68	0.06
DISEASE	278	0.24
LAB_CHECK	66	0.06
IMAGE_CHECK	38	0.03
BODY	573	0.50
MEDICINE	133	0.12

在实验中，训练样本与测试样本的比例为85%∶15%。相关参数设置如下：查找层矩阵采用基于Word2Vec的skip_gram模型[20]预训练初始化，batch_size的大小设置为64，训练轮数为60轮，学习率为0.001，查找层矩阵的向量维度与LSTM隐藏层大小设置为300，训练优化器采用Adadelta，BERT的参数均设置为默认的参数。

3.2 实验结果

以BiLSTM+CRF、BERT作为baseline，各模型在CCKS2017数据集上的识别效果如表3所示。由实验结果可知，本文所提出的方法在性能上优于当前主流的两种算法，并且其领先幅度具有显著性(显著性检验P值远小于0.05)。进而，我们研究了各种算法在不同类型实体上的识别效果，其结果如表4所示。

表3 各模型在CCKS2017数据集上的识别效果

模型	P	R	F
BiLSTM+CRF	92.95	92.64	92.79
BERT	92.83	94.10	93.46
本文的模型	93.32	94.20	93.81

表4 各模型在CCKS2017数据集各实体上的效果对比

		BiLSTM+CRF	BERT	本文模型
DISEASE	P	80.94	85.09	86.88
	R	76.11	78.23	79.23
	F	78.45	81.51	82.87
TREATMENT	P	84.33	86.63	92.17
	R	83.19	87.80	89.40
	F	83.76	87.21	90.76
SIGN	P	96.15	96.13	96.00
	R	97.89	97.90	97.79
	F	97.01	97.00	96.89
CHECK	P	96.33	95.81	96.03
	R	97.03	97.56	97.22
	F	96.68	96.68	96.62
BODY	P	89.27	88.90	87.80
	R	88.43	87.11	90.90
	F	88.84	87.99	89.32

在训练时，观察到SIGN类别的收敛速度最快，

① https://www.biendata.com/competition/ CCKS2017_2/

② https://biendata.com/competition/ccks_2019_1/

而 TREATMENT 类别和 DISEASE 类别的收敛速度最慢。而最终的各实体识别效果中，效果最好的实体类别为 SIGN 与 CHECK；DISEASE、TREATMENT、BODY 类别实体的识别效果一般。整体上看，本文的模型在中文电子病历数据集上的 F 值达到了 93.81，比 BiLSTM＋CRF 模型高出约 1.02 个百分点；比 BERT 模型高出 0.35 个百分点。从各实体类别的结果上看，除了 SIGN 和 CHECK 类别的效果略有降低外，其他类别的识别效果相比 BiLSTM＋CRF 与 BERT 模型都有不同程度的提升，显示出了本文方法的优势所在。

同时，各模型在 CCKS2019 数据集上的识别效果如表 5 所示。类似地，我们也研究了各种算法在不同类型实体上的识别效果，其结果如表 6 所示。

表 5　各模型在 CCKS2019 数据集上的识别效果

模型	P	R	F
BiLSTM＋CRF	81.90	80.66	81.28
BERT	81.92	85.00	83.43
本文的模型	83.00	85.95	84.46

表 6　各模型在 CCKS2019 数据集各实体上的效果对比

		BiLSTM＋CRF	BERT	本文模型
DISEASE	P	81.25	80.41	81.03
	R	78.41	80.44	81.26
	F	79.80	80.42	81.14
OPERATION	P	83.50	78.00	81.19
	R	78.17	77.33	79.98
	F	80.75	77.66	80.58
LAB_CHECK	P	92.11	91.32	91.25
	R	91.84	90.95	91.01
	F	91.97	91.13	91.13
IMAGE_CHECK	P	88.41	86.58	89.61
	R	85.50	88.65	87.33
	F	86.93	87.60	88.46
MEDICINE	P	94.51	90.10	93.98
	R	90.21	93.97	93.26
	F	92.31	91.99	93.61
BODY	P	77.24	79.49	83.04
	R	77.14	79.63	86.00
	F	77.19	79.56	84.49

从 CCKS2019 的结果来看，整体上本文的模型 F 值达到了 84.46，并且同样显著优于两种主流算法（显著性检验 P 值远小于 0.05）。各类别的实体中，除了 OPERATION 类与 LAB_CHECK 类实体的 F 值略有下降之外，其他类别的实体识别效果相较于 BiLSTM＋CRF 模型和 BERT 模型都有不同程度的提升。

此外，为了验证加入部首—标签矩阵对医疗实体识别效果的提升，我们还将 BERT＋CRF 与 BERT＋引入部首标签矩阵的 CRF 在两个数据集上的实验结果进行比较，实验结果如表 7、表 8 所示。

表 7　BERT＋CRF 与 BERT＋引入部首—标签矩阵的 CRF 在 CCKS2017 数据集上的效果对比

		BERT＋CRF	BERT＋引入部首—标签矩阵的 CRF
DISEASE	P	85.08	86.73
	R	78.23	79.03
	F	81.51	82.70
TREATMENT	P	87.63	92.18
	R	88.11	89.19
	F	87.87	90.66
SIGN	P	96.13	96.62
	R	98.07	98.07
	F	97.09	97.34
CHECK	P	95.81	96.25
	R	97.76	97.49
	F	96.78	96.87
BODY	P	88.90	87.76
	R	91.27	90.89
	F	90.07	89.29

表 8　BERT＋CRF 与 BERT＋引入部首—标签矩阵的 CRF 在 CCKS2019 数据集上的效果对比

		BERT＋CRF	BERT＋引入部首—标签矩阵的 CRF
DISEASE	P	77.33	79.47
	R	80.00	82.03
	F	78.56	80.72
OPERATION	P	85.14	84.29
	R	87.75	87.78
	F	86.43	86.00

续表

		BERT+CRF	BERT+引入部首—标签矩阵的CRF
LAB_CHECK	*P*	80.41	82.43
	R	86.23	88.41
	F	83.22	85.31
IMAGE_CHECK	*P*	86.21	88.29
	R	90.09	88.29
	F	88.11	88.29
MEDICINE	*P*	92.20	92.66
	R	94.37	94.84
	F	93.27	93.74
BODY	*P*	81.26	83.02
	R	84.67	85.51
	F	82.93	84.25

从BERT+CRF模型，与将CRF模型修改引入部首—标签矩阵后的BERT+CRF的结果对比来看，在CRF层加入部首—标签矩阵对大部分的实体的识别效果是有一定提升的，仅在CCKS2017数据集中的BODY实体和CCKS2019数据集中的OPERATION实体上识别效果有一定程度的下降，对于其他类别的实体均有不同程度的提升，进一步验证了CRF层引入部首—标签矩阵在医疗病历数据集上的提升效果。

3.3 案例分析

本节主要结合一些具体的案例，分析模型的提升效果。

案例一 “术后予头孢美锉钠抗感染，止血，及补液等对症支持治疗。”

在该案例中，药物类别的正确实体词为“头孢美锉钠”。而不加入部首特征的模型抽取出来的药物实体词为“头孢美锉钠抗感染，止血，及补液”；推测原因为模型习得的模式为倾向于将有“予以”“治疗/对症治疗”的上下文的词标注为药物，因此对于中间有一些非药物的治疗方案的词没有剔除能力。而加入部首特征后，可以将“头孢美锉钠”一词提取出来，体现了部首特征发挥的作用。

案例二 “于1个多月前因食管下段癌于我院行胸腹腔镜联合全胸段食管切除+食管—胃底右颈部吻合术。”

在该案例中，手术类别的正确实体词为“胸腹腔镜联合全胸段食管切除+食管—胃底右颈部吻合术。”对于传统的BiLSTM+CRF模型，识别出的实体词为“胸腹腔镜联合全胸段食管切除”，而本文的模型可以将手术实体词完整抽取出来。推测原因，“术”的部首对应的E-OPERATION的得分较高，从而使得模型倾向于将整个“术”之前的字符包括进手术实体中。

案例三 “患者3月余前因直肠癌于我院行腹腔镜直肠癌根治术(DIXON术)，术后常规病理示：(直肠)黏液腺癌伴印戒细胞癌成分(40%)，肿瘤切面积8.5*1.7cm。”

在该案例中，手术类别的正确实体词为“腹腔镜直肠癌根治术(DIXON术)”，而本文的模型提取的手术类别实体词为“腹腔镜直肠癌根治术”和“DIXON术”。推测可能的原因应该与案例二一致。这也解释了本文的OPERATION实体词的识别效果略有降低的原因：存在着一部分带缩写注释的手术名，而本文的模型倾向于将其识别为两个手术实体词。

案例四 “患者肺部轻度慢性发炎。”

在该案例中，症状类别的实体词为“轻度慢性发炎”，轻度慢性发炎这种带有多个修饰词的症状词在数据集中比较少见，传统的命名实体识别模型无法抽取出该症状实体词。本文的模型通过结合了部首信息很好地抽取出了以“炎”的部首为结尾的、“肺”的部首为开始的中间一系列字符构成的该实体词，体现了部首信息与部首—标注得分对于这种罕见症状词的抽取能力。

案例五 “侵透肌层，黏膜上层及黏膜层均见低分化浆液性腺癌浸润，脉管内见癌栓。”

在该案例中，疾病类别的正确实体词为“低分化浆液性腺癌”，其中“低分化浆液性”也属于比较长的疾病修饰词，在传统的命名实体识别模型中只能识别出“腺癌”这一实体词，而不能将相应的修饰词一起识别。经过观察可以发现，这些修饰词的部首几乎全属于部首集合*C*，说明本文的模型可以很好地将正确的实体词抽取出来。

4 总结与展望

本文通过借鉴并拓展了通用的命名实体识别模

型 BiLSTM+CRF 与 BERT，并在此基础上结合了医疗领域实体的独有特征，而在特征提取中加入部首特征，同时修改了 CRF 的得分函数计算方式以加入部首对标注的影响力。实验证实本文所提出的方法可以稳定且显著地优于当前主流的两种解决方案，同时通过分析发现对于特定类型的实体（如疾病实体、身体部位、影像检查、药物）的识别效果提升比较明显。这体现了模型的创新点对特定场景（医疗领域）下的特定实体的特殊效果。

当然，模型仍存在着一些不足，例如，为了避免部首过多而导致 CRF 层的矩阵过大训练缓慢，目前的做法是在 CRF 的部首—标签矩阵中只加入了一部分部首（大概 20 个），这就要求使用者事先对数据集有一定程度的了解，筛选出一些对标注影响较大的部首，将它们提供给模型，今后可以尝试统计属于实体词的字符的高频部首以实现这个过程的自动化。同时，模型对于部首特征不甚明显的数据集提升效果较为有限，从而使得模型的适用范围有一定的局限性。

参考文献

[1] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging [J]. arXiv preprint arXiv: 1508.01991, 2015.

[2] 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.

[3] Farmakiotou D, Karkaletsis V, Koutsias J, et al. Rule-based named entity recognition for Greek financial texts[C]//Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000), 2000: 75-78.

[4] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]//Proceedings of the 18th Intemational Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001: 282-289.

[5] Morwal S, Jahan N, Chopra D. Named entity recognition using hidden Markov model(HMM) [J]. International Journal on Natural Language Computing (IJNLC), 2012, 1(4): 15-23.

[6] Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition [J]. arXiv preprint arXiv: 1402.1128, 2014.

[7] Felix A Gers, Jürgen Schmidhuber, Fred A Cummins. Learning to forget: Continual prediction with LSTM[J]. Neural Computation, 2000, 12(10): 2451-2471.

[8] Luo L, Yang Z, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition [J]. Bioinformatics, 2017, 34(8): 1381-1388.

[9] Lin B Y, Xu F, Luo Z, et al. Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media[C]//Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017: 160-165.

[10] Chen T, Xu R, He Y, et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN [J]. Expert Systems with Applications, 2017, 72: 221-230.

[11] Zhang Xiaojun. Statistical natural language processing (second edition) [EB/OL]. www.jstor.org/stable/44113770, 2020.

[12] Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF [J]. arXiv preprint arXiv: 1603.01354, 2016.

[13] Dang T H, Le H Q, Nguyen T M, et al. D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information [J]. Bioinformatics, 2018, 34(20): 3539-3546.

[14] Wang X, Zhang Y, Ren X, et al. Cross-type biomedical named entity recognition with deep multi-task learning [J]. arXiv preprint arXiv: 1801.09851, 2018.

[15] 王尽忠. 部首的产生发展演变及其类型[J]. 中医药文化, 1990(1): 5-8.

[16] Jinming C. Categorization of Chinese character radicals and instruction of Chinese character for teaching Chinese as foreign language [J]. Sinología Hispánica, 2017, 4(1): 63-80.

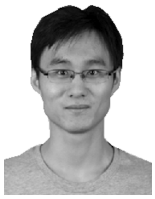
[17] 李俊红，李坤珊. 部首对于汉字认知的意义——杜克大学中文起点班学生部首认知策略测查报告[J]，世界汉语教学，2005,19(4)：4,20-32.

[18] 石民，李斌，陈小荷. 基于 CRF 的先秦汉语分词标注一体化研究 [J]. 中文信息学报，2010，24(2)：39-46.

[19] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations [J]. arXiv preprint arXiv: 1802.05365, 2018.

[20] Guthrie D, Allison B, Liu W, et al. A closer look at skip-gram modelling[C] //Proceedings of the LREC, 2006: 1222-1225.

李丹(1998—),硕士研究生,主要研究领域为自然语言处理、多模态信息检索。
E-mail: lidan528@mail.ustc.edu.cn



郑毅(1987—),博士,主要研究领域为自然语言处理、机器学习。
E-mail: zhengyi29@huawei.com

徐童(1988—),通信作者,博士,副教授,主要研究领域为数据挖掘、社交网络。
E-mail: tongxu@ustc.edu.cn

(上接第 53 页)

[14] 才智杰,孙茂松,才让卓玛. 一种基于向量模型的藏文字拼写检查方法[J]. 中文信息学报,2018,32(09):47-55.

[15] 珠杰,李天瑞,刘胜久. TSRM 藏文拼写检查算法[J]. 中文信息学报,2014,28(3): 92-98.

[16] 色差甲,贡保才让,才让加. 藏文音节拼写检查的 CNN 模型[J]. 中文信息学报, 2019, 33(01): 111-117.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.

色差甲(1991—),博士研究生,主要研究领域为藏文自然语言处理。
E-mail: sechajia@126.com

才让加(1963—),教授,博士生导师,主要研究领域为藏文自然语言处理。
E-mail: zwxxzx@163.com

慈祯嘉措(1989—),博士研究生,主要研究领域为计算语言学、机器翻译。
E-mail: 543819011@qq.com