doi:10.3969/j.issn.1673-9833.2021.05.010

基于SoftLexicon的医疗实体识别模型

张 旭[1,2]，朱艳辉[1,2]，梁文桐[1,2]，詹 飞[1,2]

（1. 湖南工业大学 计算机学院，湖南 株洲 412007；
2. 湖南工业大学 湖南省智能信息感知及处理技术重点实验室，湖南 株洲 412007）

摘 要：为了解决在中文电子病历命名实体识别任务中，基于字符粒度NER方法对序列信息遗漏的问题，以及引入外部词典资源方法所带来的运算效率问题，提出一种基于SoftLexicon的医疗实体识别模型。首先，将输入序列中的每个字符映射到一个稠密向量中；接下来，引入外部词典资源，为每个字符构造SoftLexicon特征，并将其添加到对应的字向量表示中；然后，将这些增强的字符表示放入Bi-LSTM和CRF层，以获得最终的识别结果。该模型既能有效捕捉句子序列中字符的特征，提取上下文之间的依赖关系，又能实现标签预测的顺序性。以CCKS-2020医疗命名实体识别评测任务提供的电子病历数据作为实验数据集，实验结果表明，与基于字符粒度的传统NER方法相比，所提方法在实体识别性能和效率上都显著提高。

关键词：中文电子病历；命名实体识别；SoftLexicon；BiLSTM；CRF

中图分类号：TP391.1 **文献标志码**：A **文章编号**：1673-9833(2021)05-0077-08

引文格式：张 旭，朱艳辉，梁文桐，等 . 基于SoftLexicon的医疗实体识别模型 [J]. 湖南工业大学学报，2021，35(5)：77-84.

A Medical Entity Recognition Model Based on SoftLexicon

ZHANG Xu[1,2]，ZHU Yanhui[1,2]，LIANG Wentong[1,2]，ZHAN Fei[1,2]

（1. College of Computer Science，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. Hunan Key Laboratory of Intelligent Information Perception and Processing Technology，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：In view of the problem of missing sequence information based on the character granularity NER in the task of naming entity recognition of Chinese electronic medical records, as well as the low computational efficiency brought about by the introduction of external dictionary resource methods, a model based on SoftLexicon has thus been proposed. First, each character in the sequence is mapped to a dense vector; next, an external dictionary resource is introduced to construct SoftLexicon features for each character to be added to the corresponding word vector representation; then, these enhanced characters representations are to be put into the Bi-LSTM and CRF layers so as to obtain the final recognition result. The model can effectively capture the characteristics in the sentence sequence, and extract the dependencies between contexts, thus realizing the sequentiality of label prediction. With the electronic medical record data provided by the CCKS-2020 medical named entity recognition evaluation task is as the

收稿日期：2020-12-25

基金项目：国家自然科学基金资助项目（61871432）；湖南省自然科学基金资助项目（2020JJ6089）；湖南省教育厅科研基金资助重点项目（19A133）

作者简介：张 旭（1997-），男，安徽阜阳人，湖南工业大学硕士生，主要研究方向为自然语言处理和知识工程，E-mail：Ch3ungxu@163.com

通信作者：朱艳辉（1968-），女，湖南湘潭人，湖南工业大学教授，主要从事自然语言处理和知识工程方面的教学与研究，E-mail：swayhzhu@163.com

experimental data set, the proposed method, compared with the traditional NER method based on character granularity, has significantly improved entity recognition performance and efficiency.

Keywords：Chinese electronic medical records；name entity recognition；SoftLexicon；Bi-LSTM；CRF

1 研究背景

近年来，自然语言处理技术（natural language processing，NLP）的应用越来越广泛。医疗行业信息化迅速发展，其中电子病历（electronic medical record，EMR）在临床治疗、疾病预防等方面扮演着重要角色。EMR 是医务人员在病人治疗过程中（该过程包括临床诊断、检查检验、临床治疗等），利用医疗机构信息系统生成患者的数字化信息，并进行存储、管理、传输和医疗记录的再现 [1]。对电子病历进行数据处理，构建专业且全面的医疗知识库，更有利于发挥其在“智慧医疗”中的作用。但是，目前电子病历大多处于非结构化状态，因而严重制约了其开发与利用 [2]。

命名实体识别（named entity recognition，NER）是自然语言处理技术的一个分支，属于信息抽取的子任务，它将具有特定意义的实体从非结构文本中提取出来，并将其归入预定类别，例如从文本中识别出与人名、地名和机构名相关的实体。NER 本质上可以被看成是一种序列标注问题，在许多下游任务中扮演着重要的角色，包括知识库建设 [3]、信息检索 [4] 和问答系统 [5]。

随着医疗 AI 技术的发展，信息抽取技术在医疗信息化的进程中扮演着不可或缺的角色，这一定程度上与国内外开展的相关评测任务密不可分，它们推动了大批学者对前沿技术的探索；国外的 I2B2 会议催生了一系列优秀的研究成果，HMM（hidden markov model）、CRF（conditional random field）等基于统计的机器学习方法首次被应用于医疗 NER 任务中，且有不错的性能表现；国内的全国知识图谱与语义计算大会（China Conference on Knowledge Graph and Semantic Computing，CCKS）自 2017 年起，已经连续 4a 组织中文电子病历命名实体识别相关评测。在 CCKS−2017 面向中文电子病历的命名实体识别任务中，参评者均有对 Bi-LSTM（bidirectional long short-term memory）算法模型的实现 [6]。Zhang Y. 等 [7] 分别采用 CRFs 和 BiLSTM-CRF 从电子病历数据集中识别疾病、身体部位和治疗等类型实体，对比发现后者的性能更好。CCKS−2018 评测中，何云琪等 [8] 通过结合一系列句法和语义特征表示，作为 CRF 层的输入进行标签预测；Luo L. 等 [9] 基于多特征（如标点符号、分词和词典等特征）融合，整合多种神经网络模型，完成对电子病历命名实体的识别，且取得不错的效果。潘璀然等 [10] 通过 Lattice-LSTM 网络表示句子中的单词，将字符与词序列的语义信息整合到基于字符的 LSTM-CRF 中，在 CCKS−2018 任务一上进行实验，其 F_1 值优于之前的最高结果。

但是，以上基于深度神经网络的 NER 模型，都存在不同程度的缺陷。首先，与英语 NER 相比，中文 NER 的一大难点在于中文句子不是自然地被分隔开，传统深度学习 NER 模型在中文特征提取过程中，可分为基于词粒度和基于字符粒度两大类，但由于中文电子病历实体的特殊性，即存在跨度较长的实体，因此常用分词工具无法精准识别实体边界，由此产生的分词错误会延续到上层模型的预测；基于字粒度的模型解决了分词错误的问题，但无法利用到句中单词的信息，尤其对于中文，相同字符在不同词中可能有不同的涵义，例如“灯光”和“争光”中的“光”字分别代表了“光线”和“荣誉”的含义；其次，研究者较少关注先验知识对识别效果的辅助作用 [11]，在 Zhang Y.[12] 的工作中证明了词典信息对提高 NER 准确率的重要性，但是现有引入词典的方法无一例外都建立了复杂的模型结构，导致运算效率低下，实用性不高。

综合以上问题，本文利用字符粒度 Bi-LSTM-CRF 模型的优势，提出一种基于“BMES”标签的词典简化方案，将单词词典整合到字符表示层中，SoftLexicon 方法避免了设计复杂的序列建模结构，通过对字词向量的拼接来完成词典信息引入，无需动态对句子序列进行编码，具体工作将在 2.2 节中展开介绍；同时，由于字符与词典的匹配不与 LSTM 编码层同步进行，因此很大程度上解决了引入词典带来运算效率低的问题。词典作为一种已有的先验知识，可以为字符信息提供很好的补充，增强神经网络模型对先验知识的学习，以便更完整地获取电子病历文本句中的实体特征，通过实验验证了基于 SoftLexicon 的中文电子病历实体识别模型无论在准确率还是效率上都有不错的表现。

本文后续结构如下：首先，对 SoftLexicon 方法进行概述，并对字符表示层以及序列建模层实现过

程展开介绍；然后介绍本文实验的相关工作，包括本文实验所用数据集，以及实验软硬件和参数设置，并对不同模型的对比实验效果进行分析；最后总结现有工作并提出后续工作设想。

2　基于 SoftLexicon 的中文电子病历实体识别模型

2.1　中文电子病历实体识别任务

CCKS-2020 面向中文电子病历的医疗实体抽取是 CCKS 围绕中文电子病历语义化开展的系列评测的一个延续，本文采用 CCKS-2020 评测提供的中文电子病历实体数据集，标注数据包括了医疗实体的名称、起始和结束位置以及预定义类别，其中 6 类预定义类别定义如表 1 所示。

中文电子病历命名实体识别任务要求在纯文本电子病历文档中，识别并抽取出与符合预定义类别的实体，及其在文本中的位置信息，并将它们以字典的形式表示。

表 1　CCKS-2020 预定义实体类别及定义

Table 1　CCKS-2020 predefined entity classes

实体类别	含 义 解 释
疾病和诊断	医学上定义的疾病和医生在临床工作中对病因、病生理等所作的判断。
影像检查	影像检查（X 线、CT、MR 等），造影，超声，心电图
实验室检验	临床工作中检验科进行的化验，不含免疫组化等广义实验室检查
手术	医生针对病症在病人身体局部进行切除、缝合等治疗手段，属于外科治疗方法
药物	用于疾病治疗的具体化学物质
解剖部位	指疾病、症状和体征发生的人体解剖学部位。

2.2　基于 SoftLexicon 的实体识别模型

在进行关键词自动抽取时，以 HMM、CRF 为代表的传统机器学习方法依赖人工构建大量特征工程。随着计算机硬件的快速发展，再加上医疗标注语料的逐渐完善，深度神经网络模型表现其优势，它通过模拟人类神经网络，运用多层的网络运算 [13]，能有效挖掘文本潜在语义信息，对人工难以识别的特征提取效果更好。基于 SoftLexicon 的实体识别模型如图 1 所示。

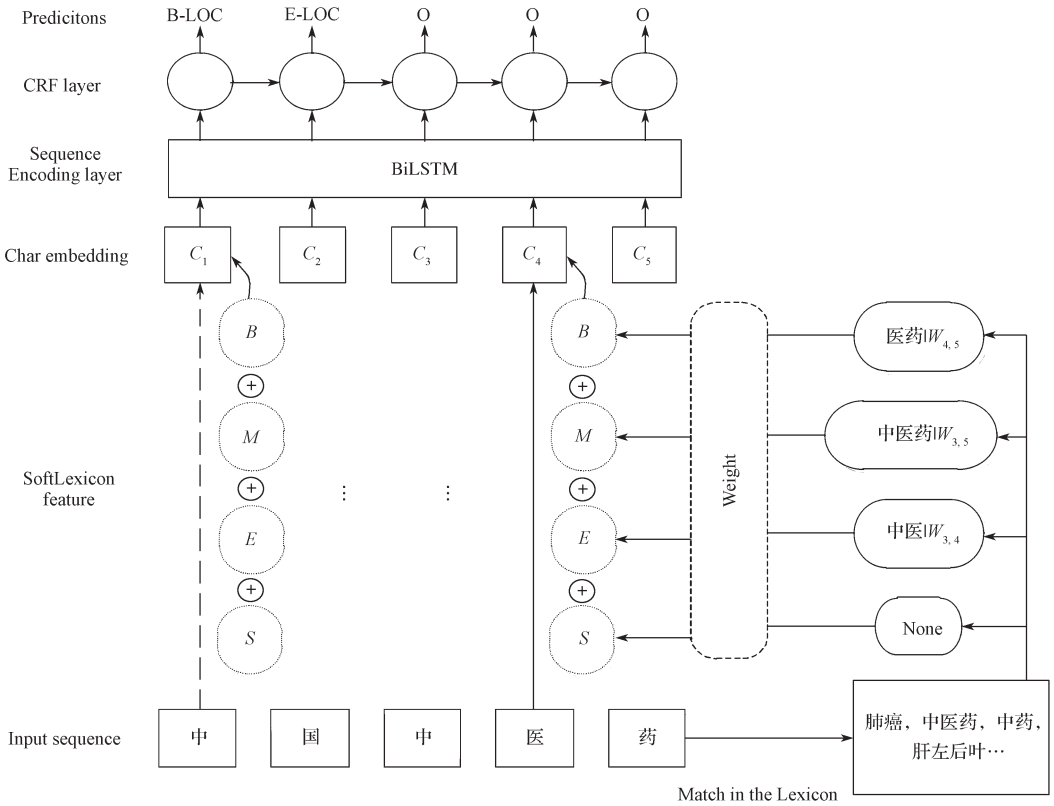

图 1　SoftLexicon 模型结构图

Fig. 1　SoftLexicon model structure

图 1 中，以输入序列“中国中医药”为例：（此图仅为流程展示，具体词典匹配结果以实验为准），整个神经网络共有 4 层结构：输入层构建输入句子的特征向量序列，分别将字符对应的 4 个单词集的表示形式组合成一个一维特征集，并将其添加到每个字符的表示形式中，例如图中的“医”字与词典匹配后

得到的相关词，进行 embedding lookup，经线性变换拼接到其字向量表示上；隐藏层为一个双向的 LSTM 网络，前向的 LSTM 用于获取前文信息，反向传播的 LSTM 用于获取下文信息，再将双向信息拼接整合；在双向 LSTM 层之上，应用 CRF（条件随机场）层为字符序列执行标签推断，CRF 能够考虑到标签之间的连续性，获得最优输出序列。

2.2.1 SoftLexicon

单纯基于字符 NER 方法的缺点是单词信息未被充分利用。考虑到这一点，Zhang Y.[12] 提出了 Lattice-LSTM 模型，用于将单词词典合并到基于字符的 NER 模型中。Lattice-LSTM 有两个优点，首先它保留了与单个字符有关的所有可能的词典匹配结果，解决了词边界不确定的问题。其次，它可以引入预训练的词向量模型，从而极大地提升了性能。然而，Lattice-LSTM 模型复杂的结构导致其运算速度十分有限。如图 2 所示，它在不相邻的字符之间额外增加了一个词级别 LSTM 通路，对字符组成的词进行编码，再输入到对应字符的 Cell 中，由此可能产生单字符对应多输入的情况，因此在模型解码阶段就增加了计算复杂度；同时 Lattice-LSTM 在引入词典过程中，依旧存在信息缺失的问题，例如图 2“中医药”中的“医”字，它只能获取到“中医”的词信息，而无法获取“医药”和“中医药”对应的词信息。

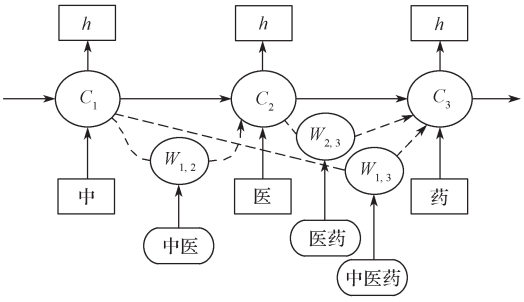

图 2 Lattice-LSTM 模型结构示意图

Fig. 2 Lattice-LSTM structure

针对上述不足，本文做了以下相关工作。课题组提出在中文电子病历 NER 上使用一种轻量级词典匹配方法，首先将输入序列 $s=\{c_1, c_2, \cdots, c_n\}$ 与词典进行匹配，得到所有相关的词 $W_{i,j}$（表示 s 子序列 $\{c_1, c_2, \cdots, c_j\}$），为了保留分段信息，将每个字符 c_i 的所有匹配单词分类为 4 个单词集“BMES”，这 4 个集合的构造如下，其中，L 表示本文所使用的词典：

$$B(c_i)=\{w_{i,k},\ \forall w_{i,k}\in L,\ i<k\leqslant n\}, \quad (1)$$

$$M(c_i)=\{w_{j,k},\ \forall w_{j,k}\in L,\ 1\leqslant j<k\leqslant n\}, \quad (2)$$

$$E(c_i)=\{w_{j,i},\ \forall w_{j,i}\in L,\ 1\leqslant j<i\}, \quad (3)$$

$$S(c_i)=\{c_i,\ \forall c_i\in L\}。 \quad (4)$$

图 3 所示为“中医药”的 Lexicon 匹配示意图。

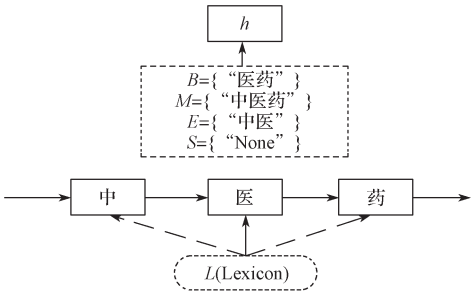

图 3 Lexicon 匹配示意图

Fig. 3 Lexicon matching

如图 3 中所示，以“中医药”为例，字符“医”与预先构造的词典进行单词匹配，得到对应的 4 个单词集：$B=\{W_{2,3}$（“医药”）$\}$，$M=\{W_{1,3}$（“中医药”）$\}$，$E=\{W_{1,2}$（“中医”）$\}$，$S=\{$（“None”）$\}$（如果没有与之匹配到的词语，就用“None”来表示该集合）。同时本文引入了预先训练好的词向量，单词集中的每个单词都会转化成对应的词向量；然后对四个单词集中的所有单词执行权重归一化，此处使用基于统计的静态加权方法 [14]，即静态数据中每个词出现的频率，这种频率能一定程度上反映该词的重要程度，静态数据可以来源于医疗领域相关的文章等，其加权方法如式（5）：

$$\boldsymbol{v}^s(S)=\frac{4}{Z}\sum_{w\in S}z(w)\boldsymbol{e}^w(w)。 \quad (5)$$

式中：S 为“BMES”单词集；

$z(w)$ 为词典中单词 w 在静态数据统计中出现的频率；

Z 为单词集中所有词出现频率之和；

$\boldsymbol{e}^w$ 为用于 embedding lookup 的词向量矩阵。

最后将 4 个单词集的表示形式组合成一个一维特征，再拼接到该字符向量的表示上，从而得到最终的输入向量。

$$\boldsymbol{e}^s(B, M, E, S)=[\boldsymbol{v}^s(B); \boldsymbol{v}^s(M); \boldsymbol{v}^s(E); \boldsymbol{v}^s(S)], \quad (6)$$

$$\boldsymbol{x}^c=[\boldsymbol{x}^c; \boldsymbol{e}^s(B, M, E, S)]。 \quad (7)$$

式中：$\boldsymbol{x}^c$ 代表字符 c 对应的字向量；$\boldsymbol{e}^s(B, M, E, S)$ 代表字符 c 匹配的单词集加权组合后的词向量。

2.2.2 LSTM 网络

RNN（recurrent neural network）模型由于可以自动保存历史信息并将其应用到当前输出中，易于捕获长距离依赖关系，这些特性十分适合处理时序信息，如序列标注问题 [15]，但是在上下文距离过长的情况下，容易产生梯度爆炸或梯度消失的问题。由此衍生而来的 LSTM，在 RNN 模型基础上增加了门控机制

和一个用于保存长距离信息的 memory cell，本文使用的 Bi-LSTM 是在单向 LSTM 的基础上，增加一层反方向的 LSTM，这样能够有效捕获某一时刻的前后文信息。

LSTM 的门控机制由输入门、遗忘门、输出门 3 部分组成。以前向 LSTM 为例，具体计算公式如下：

$$\begin{bmatrix} \boldsymbol{i}_t \\ \boldsymbol{f}_t \\ \boldsymbol{o}_t \\ \tilde{\boldsymbol{c}}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\boldsymbol{W} \begin{bmatrix} \boldsymbol{x}_t^c \\ \boldsymbol{h}_{t-1} \end{bmatrix} + b \right), \tag{8}$$

$$\boldsymbol{c}_t = \widetilde{\boldsymbol{c}}_t \odot \boldsymbol{i}_t + \boldsymbol{c}_{t-1} \odot \boldsymbol{f}_t , \tag{9}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) 。 \tag{10}$$

式（8）~（10）中：σ 为 sigmoid 函数；

⊙为按元素的点积；

$\boldsymbol{W}$ 和 b 为训练过程不断更新的参数。

前向 LSTM 与反向 LSTM 具有相同的定义，但以相反的顺序对序列进行建模。在向前和向后 LSTM 的第 i 时刻处的级联隐藏状态 $\boldsymbol{h}_i=\left[\overrightarrow{h_i};\ \overleftarrow{h_i}\right]$ 形成 c_i 的上下文相关表示。

2.2.3　CRF 模型

一个简单有效的标签模型是使用 $\boldsymbol{h}_i$ 的特性为每个输出 y_i 做出独立的标签决策。但当输出标签之间有很强的依赖性时，独立的分类决定显示出局限性。CRF 是一种基于无向图的判别式概率模型，它是指在给出一组随机输入变量的条件下，推断出另一组输出随机变量的条件概率分布模式 [15]；对于序列标注任务，CRF 输入序列为一个句子，输出序列是句中每个字符的标签，采用 CRF 可以添加对标签序列的预测约束（例如，在 B-PER 后面不能接 I-LOC），提高 NER 的识别准确率。

对于一个给定的输入序列 X，预测序列为 y，本文定义如式（11）所示的打分函数，它由两部分组成，其中，$\boldsymbol{A}$ 是转移概率矩阵，$\boldsymbol{A}_{y_i,\ y_i+1}$ 代表从 y_i 标签到 y_i+1 标签的得分；$\boldsymbol{P}$ 是经过 BiLSTM 网络输出的字符标签分数矩阵，$\boldsymbol{P}_{i,\ y_i}$ 代表第 i 个字符作为标签 y_i 的分数。

$$S(X,y)=\sum_{i=0}^{n}\boldsymbol{A}_{y_i,y_{i+1}}+\sum_{i=1}^{n}\boldsymbol{P}_{i,y_i} 。 \tag{11}$$

在训练过程中，对正确标签序列进行最大似然概率估计：

$$\log(p(y\mid X))=S(X,y)-\log\left(\sum_{\tilde{y}\in Yx}e^{s(X,\tilde{y})}\right)。 \tag{12}$$

式中：Y_X 是输入序列 X 中所有可能的标注序列。在解码阶段，利用动态规划算法，找到最高的条件概率标签序列 y*，即得分函数取得最大值对应的序列：

$$y^*=\arg\max_{\tilde{y}\in Yx}S(X,\tilde{y}) 。 \tag{13}$$

3　实验设计与结果分析

3.1　实验数据分析及预处理

本文实验的数据集来自于 CCKS-2020 的评测任务，官方提供的已标注训练数据共 1 050 条文本，为了更好地掌握数据集以便模型建模，本文对训练数据中各类别的实体数量以及长度进行了统计，具体如表 2 所示。

表 2　训练语料实体统计结果

Table 2　Entity statistics of training corpus

实体类别	数量	长度最小值	长度最大值	长度平均值
疾病和诊断	6 211	1	92	6.95
影像检查	1 490	2	15	3.86
实验室检验	1 885	1	32	4.09
解剖部位	12 660	1	125	2.48
手术	1 327	2	84	12.49
药物	2 841	2	17	3.73

从表 2 数据中可以看出，“疾病和诊断”和“解剖部位”两类实体出现最为频繁，其余各类别的实体数量分布在 1 000~3 000 个。这是由电子病历的特点所决定的，患者就医都需要进行临床诊断，检查的方式有两种，轻微病症只需药物治疗，特定疾病需手术配合药物治疗，因此药物实体总数与检查实体总数基本持平。同时，手术类实体的平均长度为 12.49，且最大实体长度达 84，这些表明了电子病历中实体的特殊性，存在许多领域词汇，因此对模型的识别准确率提出较高要求。

对于深度神经网络模型来说，1 050 条训练数据不足以满足模型对数据量的需求，本文分析训练数据后发现，数据均由多个短句组成，导致文本长度过长，且相邻短句之间语义弱关联，因此本文以“。”作为分隔符结合句末分隔，对训练数据进行拆分，最终得到 10 305 个句子序列。

同时为了验证模型训练参数效果以及结果预测效果，采用交叉验证法。如表 3 所示，本文对训练数据按照 6:2:2 的比例，将其划分为训练集、验证集和测试集。

表 3　实验数据集划分

Table 3　Experimental data division

类别	训练集	验证集	测试集
句子数	6 183	2 061	2 061

本文对评测任务两阶段中发布的医疗词典文档进行去重融合，得到一个包含 6 类实体、6 293 个医疗实体的词典，将其作为本文实验所需词典。

3.2 实验环境及参数设置

本实验基于 TensorFlow 计算框架，使用 GPU 加速，具体环境配置如表 4 所示。

表 4 实验环境配置

Table 4 Experimental environment configuration

项 目	环 境
操作系统	Windows 10(x64)
CPU	i5-9300H@2.4 GHz
GPU(显存大小)	NVIDIA GTX1650(6 GB)
内存	16 GB
固态硬盘	512 GB
Python 版本	3.6.5
Tensorflow 版本	1.14.0

本文设置字向量维数为 200，进行字词融合的词向量维度为 50；考虑模型的收敛速度，将学习率设为 0.001 5，同时，为了兼顾训练效率和后期稳定性，设置 warm up 占整个训练轮次的 0.1，0.90 的学习率指数衰减，即迭代 1 000 轮次后，学习率变为原来的 0.90；隐藏层节点数设为 300，为防止过拟合现象，Dropout 调整为 0.5，具体见表 5。经过多次实验后，验证了所设参数的合理性。

表 5 实验超参数设置

Table 5 Experimental hyperparameter setting

参数名	数值
字向量维度	200
词向量维度	50
Batch size	32
Epoch	40
LSTM 隐藏层单元数	300
Dropout	0.5
学习率	0.001 5
学习率衰减值	0.90
Warm up	0.1

3.3 评价指标

本实验评价体系包括准确率 (P)、召回率 (R) 和 F_1 值，各指标具体公式如下：

$$P=\frac{|S\cap G|}{|S|}\times 100\% \ , \tag{14}$$

$$R=\frac{|S\cap G|}{|G|}\times 100\% \ , \tag{15}$$

$$F_1=\frac{2PR}{P+R}\times 100\% 。 \tag{16}$$

式（14）~（16）中：S 为模型输出结果，记为 $S=\{S_1, S_2, \cdots, S_m\}$；

G 为人工标注结果，记为 $G=\{G_1, G_2, \cdots, G_n\}$。

用严格的等价关系确定 $S\cap G$ 为 S 和 G 的交集。当且仅当一个实体的内容、所属类别、起始下标和终止下标 4 个要素全部一致时，才认为该实体的标注结果是正确的。

3.4 实验设计与结果分析

3.4.1 模型对比实验

为验证基于 SoftLexicon 模型在中文电子病历命名实体识别上的表现，课题组设计了如下对比实验方案：

1）BiLSTM-CRF 模型。通过训练语料生成 200 维的字向量，将待预测字符序列导入 BiLSTM-CRF 中进行训练，最终得到序列预测标签。实验参数设置同表 5。

2）IDCNN-CRF 模型。基于 IDCNN（iterated dilated convolutional neural networks）的特征抽取和 CRF 的约束模型。该模型卷积核个数设置为“256，512，512”卷积膨胀率为“1，2，2”，其余实验参数设置同表 5。

3）Lattice-LSTM 模型。在 BiLSTM-CRF 基础上引入外部词典，为字符向量加入词特征，并利用门结构引导信息的流动。实验参数设置同表 5。

4）SoftLecicon 模型。在 Lattice-LSTM 基础上通过优化输入表示层编码，将字符的 4 类词典集合，结合到字符的表示中。

表 6 统计了 4 种模型在测试集上的实验表现。

表 6 模型对比实验结果

Table 6 Model performance experimental results

实体类别	BiLSTM-CRF（实验一）			IDCNN-CRF（实验二）			Lattice-LSTM（实验三）			SoftLexicon（实验四）		
	P/%	R/%	F_1/%	P/%	R/%	F_1/%	P/%	R/%	F_1/%	P/%	R/%	F_1/%
疾病和诊断	83.57	81.26	82.39	85.91	86.84	86.37	85.34	84.74	85.03	87.04	86.67	86.85
影像检查	85.54	88.42	86.95	84.06	89.12	86.51	89.19	90.32	89.75	88.23	91.78	89.96
实验室检验	90.16	86.44	88.26	91.27	88.14	89.67	91.40	88.92	90.15	93.79	90.03	91.87
手术	71.66	69.31	70.46	75.17	74.25	74.16	73.38	72.79	73.08	74.23	74.85	74.53
药物	81.29	85.21	83.20	81.59	84.11	82.83	83.12	87.66	85.32	84.34	88.06	86.15
解剖部位	80.39	84.94	82.60	81.45	83.00	82.21	85.77	83.56	84.65	88.32	85.21	86.73
综合	83.28	85.06	84.16	85.46	86.44	85.94	88.12	89.45	88.78	89.88	90.24	90.05

通过分析发现，与基于 BiLSTM-CRF 模型识别的准确率对比，在引入外部词典信息后，实验三、四所用模型在同类别实体上的识别效果表现出色，综合 F_1 值分别提升了 4.62% 和 5.89%。据分析可能是由于电子病历中实体的特殊性，单纯基于字符向量的 BiLSTM-CRF 模型不能准确定位实体的边界，导致实体识别会出现缺漏、多余的现象，这体现了引入先验词典资源的必要性。IDCNN-CRF 模型在引入卷积膨胀因子后，可以获取到长距离依赖信息，适合处理长本文句子，SoftLexicon 模型在“疾病和诊断”和“手术”类实体识别上与实验二基本持平甚至有超越。四种模型对“手术”类别实体的识别效果较差，F_1 值均低于 75.00%。分析表 2 可知，“手术”类实体总数为 1 327 个，数据量不足，导致模型参数训练效果不佳，且平均实体长度为 12.49，易产生边界预测错误的现象。此外，4 种模型均存在不同程度的识别错误问题，例如，部分相似度高的实体被错误分类、样本稀疏导致未识别出实体等。

与 Lattice-LSTM 模型识别效果对比，SoftLexicon 模型在对字符表示层进行调整后，保留了更完整的词典匹配信息，基于 SoftLexicon 的识别模型综合 F_1 值达到 90.05%，相比 Lattice-LSTM 的 F_1 值 88.78%，有 1.27% 的提升；同时，SoftLexicon 在各类实体识别效果上，P 值和 R 值比较均衡，体现了模型的稳定性。

3.4.2　模型效率对比实验

为了分析 SoftLexicon 模型在引入词典后对运算效率的影响，本文以 4 个模型在同一机器上的运行时间作为对比，结果如表 7 所示。

表 7　模型效率对比实验结果

Table 7　Model efficiency experimental comparison

项　目	BiLSTM-CRF（实验一）	IDCNN-CRF（实验二）	Lattice-LSTM（实验三）	SoftLexicon（实验四）
单个 Epoch 平均用时 /h	0.34	0.32	0.85	0.45
总运行时间 /h	6.8	6.4	8.67	5.4

实验效率上，前两个模型均迭代 20 个 Epoch，实验三和实验四引入词典的方法，为防止过拟合现象，在运行 12 个 Epoch 后提前终止了迭代；通过分析表格，实验四单个 Epoch 的平均运行时长约 0.45 h，总运行时间为 5.4 h，相比实验三的单个 Epoch 所用时长减少 0.40 h，总时长缩短约 3.2 h。引入外部词典的 NER 方法相比实验一、二的方法，不可避免地会增加运算量，但 SoftLexicon 方法在计算速度上仍有不错的表现。这可能是由于 Lattice-LSTM 在不相邻的字符之间额外增加了一个词级别 LSTM 通路，对字符组成的词进行编码，再输入到对应字符的 Cell 中，因此解码阶段需耗费大量运算时间；而 SotfLexicon 方法是通过简化词典使用，只需将整合后的字向量输入序列建模层，易于实现。

综上所述，基于 SoftLexicon 的方法无论在识别性能还是运行效率上，均有良好的表现，在中文电子病历命名实体任务上具有可行性。

4　结语

为了解决传统中文电子病历 NER 方法对字符信息遗漏以及引入外部词典资源的效率问题，本文提出了一种简单有效地整合词典信息到字符表示层中的方法，优化了字符表示层的模型结构，该方法融合了深度学习和基于词典方法两者的优势，将更完整的字符信息输入到序列建模层中，在中文电子病历 NER 评测任务中，取得了不错的效果。后续工作可从如下 3 方面改进：

1）针对中文电子病历中存在实体类别不均衡的现象，采取过采样或欠采样的方法，均衡各类别数量，以提升效果较差的实体识别效果 [11]；

2）寻找字符信息更简单且准确的特征表示；

3）BERT、ALBERT 等预训练语言模型在 NLP 多个任务中均取得不错效果，考虑引入合适的预训练语言模型。

参考文献：

[1] 中华人民共和国国家卫生和计划生育委员会 . 卫生部关于印发《电子病历基本规范 (试行)》的通知 [EB/OL]. [2020−03−04]. http：//www.nhc.gov.cn/zwgk/wtwj/201304/a99a0bae95be4a27a8 b7d883cd0bc3aa.shtml.
National Health and Family Planning Commission of the People’s Republic of China. Notice of the Ministry of Health on Printing and Distributing the Basic Standard of Electronic Medical Record (Trial)[EB/OL]. [2020−03−04]. http：//www.nhc.gov.cn/zwgk/wtwj/201304/a99a0bae95be4a27a8 b7d883cd0bc3aa.shtml.

[2] WANG Y S，WANG L W，RASTEGAR-MOJARAD M，et al. Clinical Information Extraction Applications：a Literature Review[J]. Journal of Biomedical Informatics，2018，77：34−49.

[3] RIEDEL S，YAO L，MCCALLUM A，et al. Relation Extraction with Matrix Factorization and Universal Schemas[C]//Proceedings of NAACL-HLT 2013.

Atlanta：Association for Computational Linguistics，2013：74−84.

[4] CHEN Y，XU L，KANG L ，et al. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks[C]//The 53rd Annual Meeting of the Association for Computational Linguistics(ACL2015). Beijing：ACL-IJCNLP，2015：167−176.

[5] DIEFENBACH D，LOPEZ V，SINGH K，et al. Core Techniques of Question Answering Systems over Knowledge Bases：a Survey[J]. Knowledge and Information Systems，2018，55(3)：529−569.

[6] 杨飞洪，张 宇，覃 露，等 . 中文电子病历的命名实体识别研究进展 [J]. 中国数字医学，2020，15(2)：9−12.
YANG Feihong，ZHANG Yu，QIN Lu，et al. A Research Progress of Clinical Named Entity Recognition from Chinese Electronic Medical Records[J]. China Digital Medicine，2020，15(2)：9−12.

[7] ZHANG Y，WANG X，HOU Z，et al. Clinical Named Entity Recognition from Chinese Electronic Health Records via Machine Learning Methods[J]. JMIR Medical Informatics，2018，6(4)：E50.

[8] 何云琪，刘苏文，钱龙华，等 . 基于句法和语义特征的疾病名称识别 [J]. 中国科学（信息科学），2018，48(11)：1546−1557.
HE Yunqi，LIU Suwen，QIAN Longhua，et al. Disease Name Recognition Based on Syntactic and Semantic Features[J]. Science in China(Information Sciences)，2018，48(11)：1546−1557.

[9] LUO L，LI N ，LI S ，et al. DUTIR at the CCKS-2018 Task1：A Neural Network Ensemble Approach for Chinese Clinical Named Entity Recognition[C]// Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing (CCKS-Tasks 2018). Tianjin：CCKS，2018：1−6.

[10] 潘璀然，王青华，汤步洲，等 . 基于句子级 Lattice- 长短记忆神经网络的中文电子病历命名实体识别 [J]. 第二军医大学学报，2019，40(5)：497−506.
PAN Cuiran，WANG Qinghua，TANG Buzhou，et al. Chinese Electronic Medical Record Named Entity Recognition Based on Sentence-Level Lattice-Long Short-Term Memory Neural Network[J]. Academic Journal of Second Military Medical University，2019，40(5)：497−506.

[11] 李 纲，潘荣清，毛 进，等 . 整合 BiLSTM-CRF 网络和词典资源的中文电子病历实体识别 [J]. 现代情报，2020，40(4)：3−12，58.
LI Gang，PAN Rongqing，MAO Jin，et al. Entity Recognition of Chinese Electronic Medical Records Based on BiLSTM-CRF Network and Dictionary Resources[J]. Journal of Modern Information，2020，40(4)：3−12，58.

[12] ZHANG Y，YANG J. Chinese NER Using Lattice LSTM[J/OL]. [2020−03−04]. https://www.researchgate.net/publication/325008764.

[13] 梁文桐，朱艳辉，詹 飞，等 . 基于 BERT 的医疗电子病历命名实体识别 [J]. 湖南工业大学学报，2020，34(4)：54−62.
LIANG Wentong，ZHU Yanhui，ZHAN Fei，et al. Named Entity Recognition of Electronic Medical Records Based on BERT[J]. Journal of Hunan University of Technology，2020，34(4)：54−62.

[14] MA R，PENG M，ZHANG Q，et al. Simplify the Usage of Lexicon in Chinese NER[J/OL]. [2020−03−04]. https://www.researchgate.net/publication/335233357.

[15] 陈 伟，吴友政，陈文亮，等 . 基于 BiLSTM-CRF 的关键词自动抽取 [J]. 计算机科学，2018，45(增刊 1)：91−96，113.
CHEN Wei，WU Youzheng，CHEN Wenliang，et al. Automatic Keyword Extraction Based on BiLSTM-CRF[J]. Computer Science，2018，45(S1)：91−96，113.

（责任编辑：申 剑）