基于神经网络的医药科技论文实体识别与标注研究*

赵蕊洁　佟昕瑀　刘小桦　路永和
（中山大学信息管理学院 广州 510006）

摘要：【**目的**】为提高医药实体识别的效果、实现医药新知识的挖掘和提高医药科技论文的利用率，提出一种新的实体识别模型。【**方法**】构建基于Attention-BiLSTM-CRF的医药实体识别模型，在公开数据集GENIA Term Annotation Task和BioCreative II Gene Mention Tagging上分别对模型进行测试，进而使用该模型对生物医药论文的摘要进行实体标注。【**结果**】本文提出的模型优于其他基准模型，在两个数据集上的F1值分别为81.57%和84.23%、准确率分别为92.51%和97.85%，并且在数据不平衡的情况下更有优势。【**局限**】实体标注实验数据量和应用范围较为单一。【**结论**】基于Attention-BiLSTM-CRF的医药实体识别模型可以提高实体识别效果并实现医药新知识的挖掘。

关键词：生物医药实体识别　实体标注　神经网络　注意力机制
分类号：G350
DOI：10.11925/infotech.2096-3467.2021.1414
引用本文：赵蕊洁，佟昕瑀，刘小桦等. 基于神经网络的医药科技论文实体识别与标注研究[J]. 数据分析与知识发现，2022，6(9)：100-112.(Zhao Ruijie, Tong Xinyu, Liu Xiaohua,et al. Entity Recognition and Labeling for Medical Literature Based on Neural Network[J]. Data Analysis and Knowledge Discovery, 2022, 6(9): 100-112.)

1　引　言

随着医学技术迅速发展，医药领域的专业名词数量随之飞速增长，医药数据库或医药词典的更新任务也变得更加繁重。通过人工识别科技论文中的新药物及其属性并将其加入医药数据库或医药词典，需耗费大量时间和精力，且远远跟不上医药信息的增长速度，大大降低了信息利用率。

医药实体名称较为复杂，主要分为三类：一是生物学概念相关的专有名词，如蛋白质、基因、细胞的构成和反应等；二是药物相关的专有名词，如药物的构成、药物品牌的名称等；三是临床记录相关的专有名词，如病人名、医生名、医院名、疾病状态、治疗手段等。相较于普通文本的实体识别，医药类的实体名称包含大量数字和符号，同一个实体可能拥有不同的命名方式，不同实体可能存在相同的缩写等。

多数学者对应用于实体识别的神经网络模型的改进研究主要集中在三个方面：采用新提出的特征作为输入、采用新的模型架构以及采用新的目标函数。在提出新特征方面，张海楠等[1]利用字词联合向量，而姚霖等[2]利用词边界字向量，通过字和其在词中的位置建立联系，解决中文命名实体识别中单字难以完整反映语义、词会受分词工具质量影响的问题。在采用新的模型架构方面，研究者们不同类型的神经网络及神经网络的组合验证能否提高实体识别的效果。在使用新的目标函数方面，研究进展从仅考虑神经网络的输出得分发展到综合考虑输出得分和标签之间的转移得分。

针对医药实体的特征并结合实体识别的神经网络模型研究方向，本文提出基于Attention-BiLSTM-

通讯作者（Corresponding author）：路永和（Lu Yonghe），ORCID：0000-0002-7758-9365，E-mail：luyonghe@mail.sysu.edu.cn。
*本文系广州市科技计划基金项目（项目编号：202002020036）的研究成果之一。
The work is supported by the Science and Technology Program of Guangzhou, China (Grant No. 202002020036).

CRF 的医药实体识别模型，将注意力机制引入 BiLSTM 模型中，为句子中对当前词的标签判断有重要作用的词的特征信息进行加权，以辅助当前词标签的分类。在以往研究中，为了实现同一实体的不同表述使用同一标签的目的，通常将注意力得分函数设置为词与词之间的欧氏距离。但是，仅通过词与词间的语义相似度为当前词的上下文信息加权过于片面，忽视了句子中某些特定句式或指示词对当前词的标注作用。因此，词与词之间的联系不应仅以词间距离作为评判标准，本模型采用权值矩阵将句子中的两个词联系起来，其关系特征将作为模型参数进行自动提取。

本文主要工作包括两个方面，一是构建医药实体识别模型，二是针对医药科技论文进行医药实体标注。

（1）提出 Attention-BiLSTM-CRF 医药实体识别模型。该模型将注意力机制引入句子层级的特征提取中，挖掘句子中对实体标注有关键影响的信息。将本文模型与 BiLSTM-CRF、Deep-CNN-CRF 以及 BERT 模型进行对比，通过实验证明本文模型在标签类别极度不平衡的数据上有明显优势，在实体数量远小于非实体数量的实际应用语料中更具适用性。

（2）将本文提出的实体识别模型应用到医药类科技论文的实体标注中，实现了科技论文中新知识的自动挖掘。通过后期筛选和修正生成的已标注文件可以应用于未来的医药实体关系抽取、医药词典构建、医药本体构建等一系列研究中。

2　研究现状

在医药实体识别任务中，神经网络的输入由词特征向量和字符特征向量拼接而成。词特征向量通常采用分布式词表示（Distributed Word Representation），即词嵌入。字符特征向量用于提取词的拼写特征，帮助词的语法和语义信息的表达[3]。

实体识别模型中使用较多的神经网络类型主要有两种：一是循环神经网络（Recurrent Neural Network，RNN）及其变体，如长短期记忆网络（Long Short-Term Memory，LSTM）、双向长短期记忆网络（Bi-directional Long Short-Term Memory，BiLSTM）和门控循环单元（Gated Recurrent Unit，GRU）等；二是卷积神经网络（Convolutional Neural Network，CNN）[4-8]。采用 RNN 及其变体进行实体识别，是将实体识别问题视为序列标引问题，认为一个词的实体标签类型受到整个句子信息影响，将句子信息作为词的背景信息。使用 CNN 及其变体进行实体识别，则是将实体识别问题视作词层级的多分类任务，认为词的标签受到附近一定范围内的上下文信息影响，因此，CNN 在获取词的局部特征上有明显的优势[9]。

在神经网络模型用于英文医药实体识别的研究中，有学者采用基于 CNN[10] 及其变体的医药实体识别模型，如利用堆叠式 CNN 进行疾病命名实体识别[11]，使用多标签策略（Multi-Labeling Strategy，MLS）代替 CRF 层获取邻近标签间的关系，结果显示，MLS 相较于 CRF 层更容易执行且效率更高。还有学者采用 RNN[12] 及其变体构建模型，如 LSTM[13]、GRU[14] 等。利用 RNN 在 BioCreative II GM 数据集上进行测试[15]，发现 RNN 比 CRF 模型和深度神经网络的效果更好。Liu 等[16] 采用基于 BiLSTM 的模型对临床实体和健康信息实体进行识别，结果显示，使用 CNN 或 BiLSTM 模型提取字符特征效果优于单纯使用词嵌入方法，其中，CNN 和 BiLSTM 模型实验效果相近。Sahu 等[17] 采用 CNN 提取字符特征，并比较 BiRNN、BiLSTM 和 BiGRU 拼接 CRF 层三个模型在疾病名识别和分类任务中的表现，结果显示，在仅使用字符特征时，BiLSTM-CRF 模型的效果最佳；在使用字符特征和词特征时，BiGRU-CRF 模型的效果最佳。Gridach[18] 提出以 BiLSTM 提取的字符特征和词嵌入拼接作为输入的 BiLSTM-CRF 模型，将其应用于细胞、蛋白质和基因的实体识别，并取得了良好效果；由于该模型的性能良好且稳定，其他学者也采用相似的结构进行其他类型医药实体，如药物、化学式、物种、临床概念等实体的识别[19-21]。

2018 年，谷歌提出基于 Transformer 的双向编码表示预训练（Bidirectional Encoder Representations from Transformers，BERT）模型后[22]，基于 BERT 的命名实体识别研究从未间断。Souza 等[23] 将 BERT 与 CRF 相结合，将 BERT-CRF 模型应用于葡萄牙语的命名实体识别任务，在 HAREM I 数据集上，F1 值提高了 1%~4%；随着近年来 BERT 在自然语言处理

任务上的广泛应用，Alsentzer等[24]和Lee等[25]发现BERT在专业语料库上的应用较少，因此分别训练并公开了ClinicalBERT和BioBERT预训练模型，进一步推动了医学领域命名实体识别的发展。

注意力机制最初应用于机器翻译和图像识别领域，并取得了较好的效果，因此学者们开始尝试将注意力机制引入基于神经网络的实体识别模型的不同模块中。例如，将注意力机制应用于字符特征的构建中，Lyu等[26]利用注意力模型对词的字符嵌入进行编码，生成词的字符特征，然后将字符特征和词嵌入输入BiLSTM-CRF中进行标签预测；或者，先将词嵌入和字符特征向量相结合再加入注意力机制，Rei等[27]利用注意力机制对词嵌入和经BiLSTM生成的字符特征赋予权重。在注意力机制应用于生物医药实体识别的过程中，大部分学者利用其探索句子中词与词的关系，如Luo等[28]进行化学物质实体和疾病实体识别，在BiLSTM-CRF的基础上使用注意力机制计算当前词与整个篇章中其他词的相似度，在可能属于同一实体的不同形态的词之间构建起标注的依赖关系。

随着技术的不断发展，针对医药实体识别任务的研究越来越多，且识别效果也在不断提升。大多数医药实体识别研究利用现有的公开数据集作为实验数据集，不仅可以方便地获得已经标引了实体的训练集，而且能较容易地评价实体识别效果。然而有些实体识别模型在实际应用中的效果不够理想，还需要进一步完善。

3 医药实体识别模型构建

本文提出基于Attention-BiLSTM-CRF的医药实体识别模型，根据词的标签判断句子中有重要作用的词，并对其特征信息加权以辅助词标签的分类。模型采用权值矩阵将句子中的两个词联系起来，其关系特征作为模型参数进行自动提取。该模型主要分为三个模块：在词层级特征构建模块抽取句子中词本身的特征；在句子层级特征构建模块抽取词的上下文特征；最后，经过全连接层，使用CRF进行标签预测。模型整体框架如图1所示。

3.1 基于神经网络的词层级特征构建

词层级特征即词自身的语义、语法、词形等特征，不依赖于词的上下文信息。在医药实体识别任务中，使用人工构建词层级特征时，对专业背景知识的要求较高，并且自然语言处理工具的性能也会影响特征提取的准确性。因此，本文仅使用字符特征向量和词特征向量作为词层级特征。对于句子s的词层级特征向量序列$\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，$\boldsymbol{x}_i$由词特征向量$\boldsymbol{x}_i^w$和字符特征向量$\boldsymbol{x}_i^c$拼接而成，如公式(1)所示。

$$\boldsymbol{x}_i=\boldsymbol{x}_i^w\|\boldsymbol{x}_i^c \tag{1}$$

其中，“‖”为拼接运算。

(1) 字符特征向量$\boldsymbol{x}_i^c$

本部分实验模型分别使用BiLSTM、CNN以及BiLSTM-CNN自动构建词的字符特征向量，并比较三种模型所构建的字符特征向量对医药实体识别的影响，选择结果最优模型作为本实验的字符特征向量。

在利用神经网络模型获取词的字符特征向量前，需要先将字符映射为字符嵌入作为模型的输入。对于词w，其字符序列为$\{char_1,char_2,\cdots,char_m\}$，随机生成的字符嵌入表为$W^{ce}\in\mathbb{R}^{d^{ce}\times m}$。其中，$d^{ce}$为字符嵌入的维度，$m$为数据集中的总字符数。字符嵌入$r_i$的计算方法如公式(2)所示。

$$r_i=W^{ce}\boldsymbol{v}_{char_i} \tag{2}$$

其中，W^{ce}为字符嵌入表，$\boldsymbol{v}_{char_i}\in\mathbb{R}^{d^{ce}}$，该向量为除$char_i$对应的索引号所在维度的值为1、其他维度的值均为0的向量。字符特征向量模型的输入如公式(3)所示。

$$r=r_1\|r_2\|\cdots\|r_m \tag{3}$$

①char-BiLSTM模型生成的字符特征向量

前向LSTM接受按顺序输入的字符嵌入序列，在时刻m获得的输出为$\overrightarrow{\boldsymbol{h}}_m$，后向LSTM接受按逆序输入的字符嵌入序列，在时刻1获得的输出为$\overleftarrow{\boldsymbol{h}}_1$。因此，char-BiLSTM模型获得的字符特征向量由两个方向的输出拼接而成[29]，如公式(4)所示。

$$\boldsymbol{x}_i^c=\overrightarrow{\boldsymbol{h}}_m\|\overleftarrow{\boldsymbol{h}}_1 \tag{4}$$

②char-CNN模型生成的字符特征向量

将词的字符嵌入序列r输入CNN中，通过卷积层和池化层得到字符特征向量，最终将k个卷积核的输出拼接起来，再放入全连接层形成字符特征向

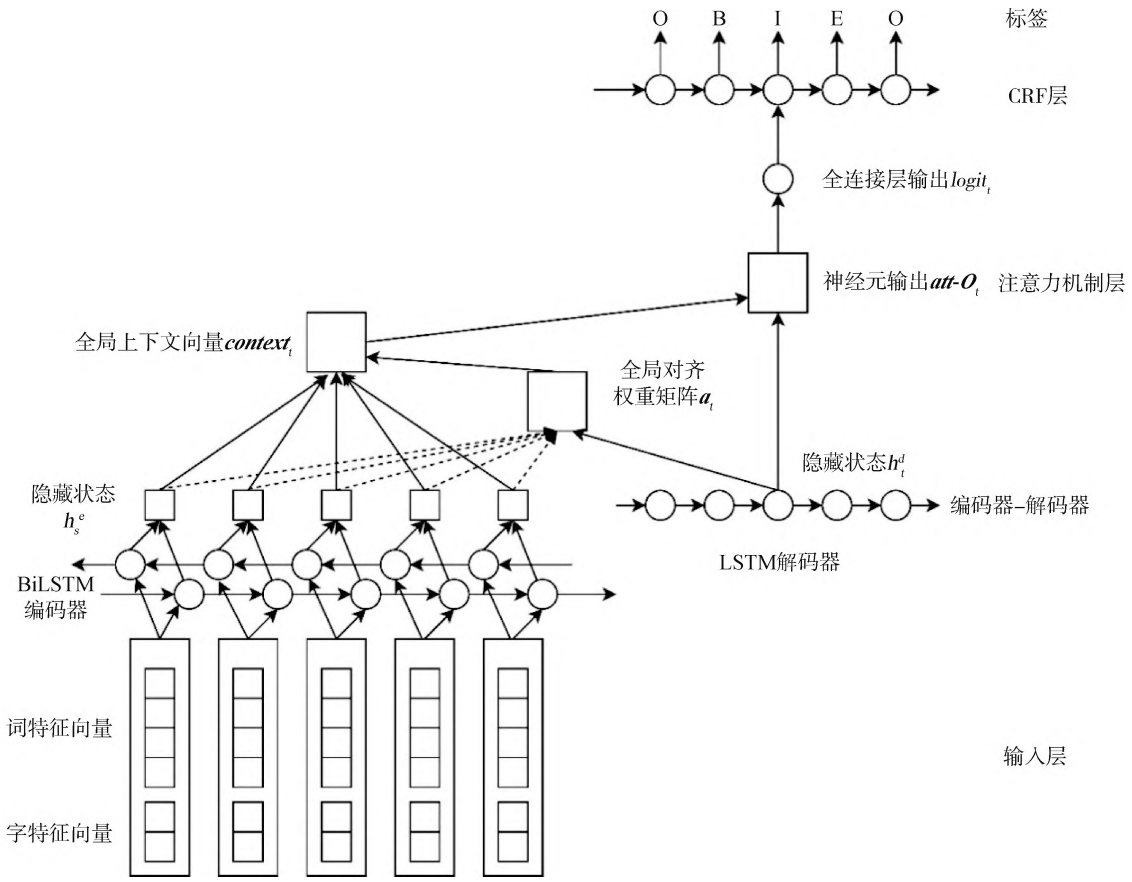

图 1　基于 Attention-BiLSTM-CRF 的医药实体识别模型框架

Fig.1　Medical Entity Recognition Model Framework Based on Attention-BiLSTM-CRF

量[10]，如公式(5)所示。

$$\boldsymbol{x}_i^c = W^c(z_{max}^{fc(1)} \| z_{max}^{fc(2)} \| \cdots \| z_{max}^{fc(k)}) \tag{5}$$

其中，W^c 是待训练的参数。

③ char-BiLSTM-CNN 模型生成的字符特征向量

由于 BiLSTM 和 CNN 在获取字符特征时的侧重点不同，因此考虑将两种字符特征融合起来，以捕获词的字符全局特征和局部特征。由于使用多窗口高度的卷积核的 CNN 和 BiLSTM 相结合可能导致整体模型参数过多，部分参数无法得到有效训练，进而降低模型的效果，因此仅使用单一窗口高度的卷积核的 CNN 和 BiLSTM 进行结合。

使用 char-BiLSTM 模型获得字符特征向量 $\boldsymbol{x}_i^{bc}$，使用 char-CNN 模型获得字符特征向量 $\boldsymbol{x}_i^{cc}$。由于将两个特征向量直接拼接并不足以利用两种不同特征之间的相互影响关系，因此选择将 $\boldsymbol{x}_i^{bc}$ 和 $\boldsymbol{x}_i^{cc}$ 分别输入全连接层中映射为维度相等的向量，并进行逐点加操作，最终形成字符特征向量[15]如公式(6)所示。

$$\boldsymbol{x}_i^c = W^{bc}\boldsymbol{x}_i^{bc} \oplus W^{cc}\boldsymbol{x}_i^{cc} \tag{6}$$

其中，$\oplus$ 为逐点加操作，W^{bc} 和 W^{cc} 为待训练的参数。

(2) 词特征向量

由于医药领域文本中包含较多的专业词汇，通用领域的词向量表并不适用，因此采用对生僻词预测效果更佳的 Skip-Gram 模型训练无标注文本，以适用于医药领域的词向量。为解决语料数据规模巨大和医药类实体在语料中出现的频次不高的问题，在 Skip-Gram 模型的输出层使用层次 Softmax (Hierarchical Softmax)作为输出分类的优化算法。

对于词特征向量 $\boldsymbol{x}_i^w$ 的提取：使用 Skip-Gram 模型生成词嵌入作为词特征向量[30]，词向量维度设为 300，其目标函数如公式(7)所示。

$$F = \sum_{w \in C} \log p(context(w)|w) \tag{7}$$

其中，w为输入的词，$context(w)$为w的上下文词，c为语料库中词的集合。

Skip-Gram生成的词嵌入表为$W^{we} \in \mathbb{R}^{d^{we} \times N^{we}}$，其中，$d^{we}$为词嵌入的维度，$N^{we}$为词嵌入表中包含的词数。对于句子$s$的词序列$\{word_1, word_2, \cdots, word_n\}$，词特征向量的计算方法如公式(8)所示。

$$\boldsymbol{x}_i^w = W^{we} \boldsymbol{v}_{word_i} \tag{8}$$

其中，W^{we}为词嵌入表，$word_i$对应的索引号所在维度的值为1、其他维度的值均为0，其向量表示为$\boldsymbol{v}_{word_i} \in \mathbb{R}^{d^{we}}$。

3.2 基于Attention-BiLSTM的句子层级特征构建

注意力机制(Attention)其本质特征为通过模仿人类的注意力方法，通过注意力分配参数筛选出特定的信息，其主要作用是衡量特征权重。也可以理解为对于某一时刻的输出，在输入层给予各个部分的注意力是权重，即输入的各个部分对于某时刻输出贡献的权重。

由于大部分医药实体包含医药领域的通用词以及非医药领域的通用词，仅依靠词层级的特征判断词的实体标签容易产生误判，需要添加隐含上下文信息的句子层级特征，以提高实体识别准确率。因此，将注意力机制引入BiLSTM模型，构建Attention-BiLSTM模型获取词的上下文信息。模型分为编码器和解码器两部分。

(1) 编码器BiLSTM模型

本文参考Luong等[4]提出的计算公式，将词层级特征向量序列$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$输入前向LSTM和后向LSTM中，编码器每一时刻的输出$\boldsymbol{h}_t^e$由前向LSTM和后向LSTM在该时刻的隐藏状态拼接而成，如公式(9)所示。

$$\boldsymbol{h}_t^e = \overrightarrow{\boldsymbol{h}_t^e} \| \overleftarrow{\boldsymbol{h}_t^e} \tag{9}$$

(2) 解码器LSTM模型

①将上一时刻$t-1$的输出$\boldsymbol{att_o}_{t-1}$和编码器t时刻的隐藏状态$\boldsymbol{h}_t^e$进行拼接，得到每个时刻t神经元的输入$\boldsymbol{input}_t^d$，如公式(10)所示。

$$\boldsymbol{input}_t^d = \boldsymbol{att\_o}_{t-1} \| \boldsymbol{h}_t^e \tag{10}$$

②将t神经元的输入传入LSTM获得每个时刻t的隐藏状态$\boldsymbol{h}_t^d$。对于固定时刻t，为获得其他时刻的状态对当前时刻的预测的重要程度，计算编码器所有时刻的隐藏状态与解码器时刻t的隐藏状态的得分。因此，定义时刻t的解码器隐藏状态$\boldsymbol{h}_t^d$和时刻s的编码器隐藏状态$\boldsymbol{h}_s^e$的得分如公式(11)所示。

$$score(\boldsymbol{h}_t^d, \boldsymbol{h}_s^e) = \boldsymbol{h}_t^{d^T} W_m \boldsymbol{h}_s^e \tag{11}$$

其中，W_m为需要训练的参数，用于自动获取$\boldsymbol{h}_t^d$和$\boldsymbol{h}_s^e$的关系特征。

③使用Softmax函数对得分进行归一化，生成全局对齐权重矩阵$\boldsymbol{a}_t$。矩阵中每一项$a_t(s)$的计算方法如公式(12)所示。

$$\begin{aligned} a_t(s) &= alignment(\boldsymbol{h}_t^d, \boldsymbol{h}_s^e) \\ &= \frac{\exp(score(\boldsymbol{h}_t^d, \boldsymbol{h}_s^e))}{\sum_{\tilde{\boldsymbol{h}}_s^e \in \boldsymbol{h}^e} \exp(score(\boldsymbol{h}_t^d, \tilde{\boldsymbol{h}}_s^e))} \end{aligned} \tag{12}$$

④对每一时刻的编码器隐藏状态赋权值，使得模型能重点关注对预测有关键作用的位置信息。使用归一化的得分$\boldsymbol{a}_t = \{a_t(1), a_t(2), \cdots, a_t(n)\}$和编码器隐藏状态序列$\boldsymbol{h}^e = \{h_1^e, h_2^e, \cdots, h_n^e\}$求得时刻$t$的全局上下文向量$\boldsymbol{context}_t$，如公式(13)所示。

$$\boldsymbol{context}_t = \boldsymbol{a}_t \otimes \boldsymbol{h}^e \tag{13}$$

其中，⊗为逐点乘操作。

⑤为综合考虑全局上下文信息和当前时刻t的信息，将$\boldsymbol{context}_t$和解码器隐藏状态$\boldsymbol{h}_t^d$拼接起来输入一个全连接层，最终获得引入注意力机制的LSTM神经元输出$\boldsymbol{att_o}_t$，如公式(14)所示。

$$\boldsymbol{att\_o}_t = W_a(\boldsymbol{context}_t \| \boldsymbol{h}_t^d) \tag{14}$$

其中，W_a为待训练的参数。

⑥$\boldsymbol{att_o}_t$经过一个全连接层转换为一个1维、时刻t的词的实体标签得分向量，如公式(15)所示。

$$\text{logit}_t = W_s \boldsymbol{att\_o}_t \tag{15}$$

其中，W_s为待训练的参数。

3.3 基于CRF的标签预测

实体标签除了与其对应的词的信息有关外，还与邻近的标签存在联系，比如实体结束标签E的后一个标签不能为实体内部标签I，非实体标签O后的下一个标签不能为实体内部标签I等。由于CRF能同时考虑模型输出和标签之间的得分以及标签之间的转移得分，本文选用此标签预测模型用于实体识别的标签预测模块。

设标签序列为$y = \{y_1, y_2, \cdots, y_n\}$，该序列的得

分如公式(16)和公式(17)所示。

$$score(x,y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n-1}A_{y_i,y_{i+1}} \tag{16}$$

$$P_{i,j}=\frac{\exp(\text{logit}_{i,j})}{\sum_{j\in[1,l]}\exp(\text{logit}_{i,j})} \tag{17}$$

其中，$\boldsymbol{P}\in\mathbb{R}^{n\times k}$为全连接层输出的得分矩阵，$P_{i,j}$表示句子中第$i$个词被标注为第$j$个标签的得分。$\boldsymbol{A}\in\mathbb{R}^{l\times l}$为标签的转移矩阵，$l$为标签数，$A_{i,j}$表示第$i$个标签转移到第$j$个标签的概率。

对分数进行归一化，求得该标签序列的概率值，如公式(18)所示。

$$p(y|x)=\frac{\exp(score(x,y))}{\sum_{\tilde{y}\in y_x}\exp(score(x,\tilde{y}))} \tag{18}$$

在模型训练阶段需最大化正确标签序列的对数似然，以获得参数的最优解，如公式(19)所示。

$$\log(p(y|x))=score(x,y)-\log\left(\sum_{\tilde{y}\in y_x}\exp(score(x,\tilde{y}))\right) \tag{19}$$

在测试阶段，取使得分最高的标签序列作为结果输出，如公式(20)所示。

$$y=argmax_{\tilde{y}\in y_x}(score(x,\tilde{y})) \tag{20}$$

4　医药实体识别实验及结果分析

4.1　实验过程

实体识别实验细分为两组实验：第一组实验通过对比 char-BiLSTM、char-CNN 和 char-BiLSTM-CNN 三种模型生成的字符特征向量对医药实体识别的影响，以得到效果最佳的字符特征向量表示模型；第二组实验采用效果最佳的字符特征向量表示模型，对比 Attention-BiLSTM-CRF 模型、BiLSTM-CRF 模型、Deep CNN-CRF 模型和 BERT 模型在不同数据集上的命名实体识别表现。

(1) 数据集的选择及预处理

本文选择两个公开的英文生物医药实体识别数据集 GENIA Term Annotation Task(简称 GENIA 数据集)和 BioCreative II Gene Mention Tagging(简称 BCII-GM 数据集)作为实验对象，数据集的语料大多数来自 Medline 数据库、DrugBank 数据库等生物医药领域的文献或摘要。其中，GENIA 数据集由 1 999 条 Medline 的论文摘要组成，主题为分子生物学，包含蛋白质、基因、细胞等实体名；BCII-GM 数据集由 Medline 随机抽取的句子组成，主题为生物化学、分子生物学和遗传学，包含基因、基因产物等实体名。数据处理步骤如下：

①将实验数据集的数据进行格式统一，采用 BIOES 实体标注方案对数据进行标注和预处理，构建词嵌入表和字符嵌入表，词嵌入表以无标签的数据集为语料进行训练，字符嵌入表则是随机生成；

②对每个批内的数据进行填充，使每个批中的每个词、每个句子的长度统一；

③对每个批内的数据进行编码，每个词的编码由词编码和字符编码组成。

两个数据集的标签类别均存在不平衡问题，非实体占比远高于实体，特别是 BCII-GM 数据集，这种情况与实际情况相近。在 GENIA 数据集中，多词实体约为单词实体的两倍，而在 BCII-GM 数据集中，单词实体略多于多词实体。从句子层面看，GENIA 数据集每一个句子都包含一个以上实体，而 BCII-GM 数据集中有大部分句子不包含实体，更接近现实语料的情况。对两个数据集的标签类型进行统计，结果如表 1 所示。

因此，数据集被划分为训练集、验证集和测试集，具体如表 2 所示。

表 1　GENIA 数据集和 BCII-GM 数据集的标签数

Table 1　Number of Tags in GENIA and BCII-GM

	标签名	GENIA		BCII-GM	
		标签数量	标签占比	标签数量	标签占比
训练集	标签 B	27 110	9.29%	8 012	2.08%
	标签 I	20 089	6.89%	6 294	1.63%
	标签 E	27 110	9.29%	8 012	2.08%
	标签 S	15 958	5.47%	10 633	2.76%
	标签 O	201 456	69.06%	352 224	91.45%
验证集	标签 B	8 972	9.19%	1 321	2.03%
	标签 I	6 967	7.14%	1 061	1.63%
	标签 E	8 972	9.19%	1 321	2.03%
	标签 S	5 008	5.13%	1 889	2.91%
	标签 O	67 696	69.35%	59 432	91.40%
测试集	标签 B	9 475	9.76%	1 394	2.14%
	标签 I	6 935	7.14%	1 035	1.59%
	标签 E	9 475	9.76%	1 394	2.14%
	标签 S	4 684	4.82%	1 882	2.88%
	标签 O	66 522	68.52%	59 544	91.26%

表2　GENIA数据集和BCII-GM数据集的句子数与词数

Table 2　Number of Sentences and Words in GENIA and BCII-GM

	GENIA		BCII-GM	
	句子数	词数	句子数	词数
训练集	11 127	15 294	14 975	36 063
验证集	3 709	18 081	2 500	39 827
测试集	3 710	20 555	2 500	43 304

（2）参数设置

词嵌入表以无标签的数据集作为语料，使用Gensim工具包中的Word2Vec API进行训练。词嵌入维度设为300，采用Skip-Gram模型，使用层次Softmax策略进行求解。字符嵌入表随机生成，字符嵌入维度设为100。

第一组对比实验中，字符特征提取实验参数如表3所示。

表3　字符特征提取实验参数设置

Table 3　Parameter Setting of Character Feature Extraction

模型	参数值
BiLSTM	隐层神经元数为100
CNN	卷积核高度为2、4、5 卷积核数量均为50 全连接层神经元数为100
BiLSTM-CNN	BiLSTM隐层神经元数为100 CNN卷积核高度为4 卷积核数为100

第二组对比实验中，Attention-BiLSTM-CRF、BiLSTM-CRF、Deep CNN-CRF模型均采用在第一组对比实验中表现最佳的生成字符特征向量的模型。句子层级特征提取实验参数如表4所示。

表4　句子层级特征提取实验参数设置

Table 4　Parameter Setting of Sentence Level Feature Extraction

模型	参数值
Attention-BiLSTM-CRF	编码器神经元数为100 解码器神经元数为200 全连接神经元数为100
BiLSTM-CRF	隐层神经元数为100
Deep CNN-CRF	卷积核高度为5 卷积核数为100 网络层数为6

实验所采用的模型参数设置如表5所示。

表5　实验模型参数设置

Table 5　Paramaters Setting

模型 参数	Attention-BiLSTM-CRF/ BiLSTM-CRF/ Deep CNN-CRF	BERT
批次大小	70	10
最大训练次数	80	5
随机失活概率	0.5	0.1
学习率	0.01	0.0001
学习率预热步数	–	500
学习率衰减率	0.9	–
优化算法	Adam	Adam
早停法参数	10	–

4.2　实验结果与分析

实验分别验证不同字符特征向量模型对实体识别的影响以及对比本文模型和现有的两种模型的实体识别效果，评价指标采用最常用的指标，即准确率、F1值、精准率和召回率，其中，精准率和召回率作为模型整体效果的辅助指标，决定性指标取决于准确率和F1值。准确率（Accuracy）用于词的层级评价模型的效果，F1值（F1 Score）用于实体的层级评价模型的效果。

准确率的计算方法如公式（21）所示。

$$accuracy = \frac{pred\_correct\_words}{totol\_words} \tag{21}$$

其中，*pred_correct_words*为词的标注标签与实际标签相同的词的总数，*total_words*为数据集中所有词的总数。

在模型标注准确的实体数（*pred_correct_entities*）大于0时，*F*1值的计算方法如公式（22）-公式（24）所示。

$$F1 = \frac{2pr}{p + r} \tag{22}$$

$$p = \frac{pred\_correct\_entities}{total\_pred\_entities} \tag{23}$$

$$r = \frac{pred\_correct\_entities}{total\_correct\_entities} \tag{24}$$

若模型标注准确的实体数为0，则p、r和$F1$值均为0。

模型对测试集的标注以词为单位，因此需用标签序列抽取出实体。实体由一个三元组（起始标签、

实体起始位置和实体长度)表示,当模型标注的实体三元组与实际标注的实体三元组完全相同时才视为实体标注准确。理论上,BIOES方法在实体标注过程中,实体对应的标签序列形式应为[B,I,…,I,E]、[B,E]或[S],即B为实体起始位置,I为实体内部,O为非实体,E为实体结束位置,S为词本身只有一个实体。

由于模型的标注会存在误差,针对特殊标签序列的情况按以下策略进行实体抽取,具体抽取策略如表6所示。

①当匹配到B或S时,视为新实体的开始位置,如果前一标签为B或I,也被视作前一实体的结束位置。

②当匹配到O时,前一标签即使为B或I,但缺少E作为结束标签,前一标签也被视作实体的结束位置。

③当标签O、E、S的后一标签为I或E时,即使缺少B作为开始标签,后一标签也被视作新实体的开始位置。

表6　特殊标签序列的部分实体抽取策略

Table 6　Entity Extraction Strategies for Special Tag Sequences

标签序列	策略
…,B,I,…,I,B(S),…	将[B,I…,I]视为模型标注的实体
…,B,I,…,I,O,…	将[B,I…,I]视为模型标注的实体
…O,B(I\E),O,…	将[B(I\E)]视为模型标注的实体
…,O,I(E),I,…,I,E(I),O,…	将[I(E),I,…,I,E(I)]视为模型标注的实体
…,S(E),I(E),I,…,I,E(I),O,…	将[I(E),I,…,I,E(I)]视为模型标注的实体

(1) 字符特征向量模型对比实验结果

不同的字符向量对比实验结果如表7所示。

表7　字符特征向量模型在数据集上的测试结果

Table7　Test Results of Character Feature Vector Model

模型 \ 数据集	GENIA				BCII-GM			
	精准率	召回率	F1	准确率	精准率	召回率	F1	准确率
char-BiLSTM+Attention-BiLSTM-CRF	81.73%	81.42%	**81.57%**	**92.51%**	**84.40%**	**84.07%**	**84.23%**	**97.85%**
char-CNN+Attention-BiLSTM-CRF	**81.79%**	78.18%	79.95%	91.96%	81.22%	77.87%	79.51%	97.15%
char-BiLSTM-CNN+Attention-BiLSTM-CRF	80.69%	**82.20%**	81.44%	92.43%	83.44%	83.49%	83.46%	97.72%

在GENIA数据集上,三种字符特征向量对实体识别效果影响不大,在实际应用中处理更大量的数据时模型间的效果差异可能会被放大。char-BiLSTM模型在F1值和准确率上表现最好,分别为81.57%和92.51%;char-CNN模型精准率最高,为81.79%,说明CNN能较好地挖掘实体词的特征,更精确地识别出实体,但其抽取特征范围较窄,对于实体特征不明显的实体词会出现遗漏,影响识别效果;char-BiLSTM-CNN召回率为82.20%,说明模型对于特征不明显的实体也能很好地识别,但引发的误判率也较高。三种模型在准确率方面相差较小,说明三种模型生成的字符特征向量对识别结果的影响主要在于对实体内部标签的识别,对于实体的开始和结束标签并没有准确识别,从而造成结果误差。

在BCII-GM数据集上,由于数据集中实体分布不均匀等情况,三种字符特征向量对医药实体识别的影响差别也更为显著。使用char-BiLSTM模型的综合表现最优,在精准率和召回率上均占优势。原因可能是CNN模型抽取局部字符特征的效果不能满足全文抽取实体的需求,从而影响char-CNN模型和char-BiLSTM-CNN模型的识别效果。

总体而言,char-BiLSTM模型的F1值在两个数据集上表现最好。原因主要有如下:

①相较于使用BiLSTM抽取字符特征,增加CNN会导致训练参数增加,使得整个神经网络训练难度增加,由于部分参数没有得到有效训练,因此影响实体的标注效果。

②在使用BiLSTM抽取到的信息中加入CNN抽取结果并没有增加有用的特征信息反而增加了噪音,尤其在BCII-GM数据集标签级别极度不平衡的

情况下，导致标注结果不准确。char-BiLSTM生成的字符特征提高了实体识别的总体效果。

③char-CNN生成的字符特征会使识别的召回率偏低，严重影响总体效果。

④char-BiLSTM-CNN生成的字符特征并没有融合两个模型的优点，在不同数据集上的精准率和召回率相较于另外两个模型表现不一致，但总体效果处于两个模型之间。

综上所述，char-BiLSTM的Attention-BiLSTM-CRF模型在医药实体识别上的综合表现最佳，因此对比实验采用char-BiLSTM提取字符特征向量。

（2）与基准模型对比实验结果与分析

本文模型与两个基准模型的对比实验结果如表8所示。

表8　本模型与基准模型在GENIA和BCII-GM数据集上的测试结果

Table 8　Test Results of This Model and Benchmark Model on GENIA and BCII-GM

数据集 / 模型	GENIA		BCII-GM	
	F1	准确率	F1	准确率
char-BiLSTM+BiLSTM-CRF	81.55%	**92.53%**	83.68%	97.81%
char-BiLSTM+Deep CNN-CRF	70.77%	87.85%	59.81%	94.85%
char-BiLSTM+Attention-BiLSTM-CRF	81.57%	92.51%	84.23%	**97.85%**
BERT-CRF	**84.45%**	91.99%	81.99%	97.56%
KoBioLM	–	–	**85.10%**	–
Triaffine+ BioBERT	81.23%	–	–	–

从测试集结果的评价指标方面分析，F1值差别较大，说明模型标注的差别主要在实体边界词的判断上；Deep CNN-CRF模型在两个数据集中均无法获得较好的成绩，说明在标签不平衡的情况下，仅考虑词的局部上下文信息并不足以准确判断词的标签，而使用BiLSTM的两个模型能更好地利用词的全局上下文信息辅助标签分类；将注意力机制引入BiLSTM，重新分配对整个句子的关注度，在语料中的实体分布不均匀的情况下，对注意力的重新分配有利于判断标签类别。

从实验结果来看，虽然Attention-BiLSTM-CRF模型在GENIA数据集上F1值低于BERT模型，但是在实体比例极小、数据分布极度不均衡的BCII-GM数据集上，Attention-BiLSTM-CRF模型的F1值和准确率均优于BERT模型。专门为医学实体识别设计的模型KoBioLM在BCII-GM数据集上的F1值为85.10%[31]，与本文模型F1值84.23%仅相差0.87个百分点，但KoBioLM专门为医学命名实体识别设计，不适合用于其他领域的实体识别，应用广泛性不及本模型。在GENIA数据集上，根据Paper With Code提供的2021年效果最好的模型Triaffine+BioBERT的F1值为81.23%[32]，低于Attention-BiLSTM-CRF模型的F1值。Attention-BiLSTM-CRF模型在所有实验中准确率较高，且F1值与其他模型相差不大，因此，选取Attention-BiLSTM-CRF模型进行下一步的实体标注实验。

5　医药科技论文实体标注实验及结果分析

5.1　实验过程

实体标注实验选择遗传学领域的*Genome Medicine*期刊作为数据集，该期刊与模型训练阶段所用的数据集主题相近。在*Genome Medicine*期刊中以“gene”作为关键词进行搜索，共获得994篇摘要。本实验在8核64GB内存的服务器上运行，操作系统为Windows7，编程软件为PyCharm，编程语言为Python，使用TensorFlow进行模型构建。

实体标注实验包括三个主要步骤：对获取的994篇摘要进行预处理；使用经过两个数据集训练的模型对语料进行医药实体标注；对标注结果进行修正，得到医药实体标注结果和抽取出的实体文件。

在医药实体标注中，使用Attention-BiLSTM-CRF模型对科技论文摘要语料进行实体识别，模型的原始输出包括两个文件：一是候选标注文件，二是候选实体文件。候选标注文件存储着科技论文摘要

语料的候选医药实体标注结果，数据格式为一行一个词及其标签，句子间以换行符间隔。候选实体文件存储着在语料中以[B，I，…，I，E]、[B，E]、[S]、[B，I，…，I]、[I，…，I(E)]、[B(I\E)]为标签序列的词序列组成的候选实体，数据格式为一行一个候选实体。

实体标注后需要对所得实体进行筛选，筛选规则通过对训练集数据进行统计分析得到。

设一个词 w 为训练集中的实体，定义词 w 在训练集中作为实体构成词的概率为 P^{w-e}，如公式（25）所示。

$$P^{w-e}=\frac{freq^{w-e}}{freq^{w}} \tag{25}$$

其中，$freq^{w-e}$ 表示词 w 在训练集中作为实体构成词的出现频率，即标签为 B、I、E、S 的频率，$freq^{w}$ 表示词 w 在训练集中总的出现频率，在统计时所有词均转换为小写。在两个训练集中所有实体构成词的 P^{w-e} 大部分集中在 0.9~1.0，即若一个词为构成某一实体的词，则其很大概率也是构成其他实体的词，反之则不然。设实体 e 由 n 个实体构成词组成，即 $[w_1, w_2, \cdots, w_n]$，即实体 e 内实体构成词 w_i 的 P_i^{w-e} 的平均值为 P^{ent-e}，如公式（26）所示。

$$P^{ent-e}=\sum_{i=0}^{n}P_i^{w-e}/n \tag{26}$$

对于一个候选实体，可以通过其实体构成词的 P_i^{w-e} 的平均值来粗略判断其为真实实体的概率，实验证明 90% 以上实体的 P^{ent-e} 大于 0.5，因此，使用此方法对模型在验证集上抽取出的候选实体进行简单筛选是可行的。实验模型最终输出三个文件，分别是语料标注文件、XML 格式语料标注文件以及实体文件。

5.2　实验结果与分析

从候选标注文件和实体文件可得，使用 GENIA 数据集所训练的模型（简称模型 G）共获得 15 710 个候选实体，使用 BCII-GM 数据集所训练的模型（简称模型 B）共获得 1 597 个候选实体。

经过筛选后，模型 G 共获得 12 851 个实体，模型 B 共获得 1 237 个实体。从数量上看，模型 G 标注的医药实体数量远远超过模型 B，与两个训练集的医药实体概率分布相似。两个模型抽取出的部分候选医药实体如表 9 所示。

表 9　模型 G 和模型 B 抽取的部分候选医药实体

Table 9　Candidate Pharmaceutical Entities Extracted by Model G and Model B

模型 G		模型 B	
实体名	频率	实体名	频率
RUNX3（RUNX3 基因）	1	KRAS mutant（KRAS 突变体）	7
DNA methylation（DNA 甲基化）	63	BAP1（BRCA 相关蛋白）	4
human leukocyte antigen（人类白细胞抗原）	13	Myc（Myc 癌基因组）	4
mtDNA genome（线粒体基因组）	1	β 2AR（β 肾上腺素能受体）	2
rheumatoid arthritis（类风湿性关节炎）	16	ABCG2（ABCG2 基因）	1

从抽取得到的候选实体类型可以看出，模型 G 抽取的实体类型更丰富，主要包括蛋白质、细胞、基因、疾病等类型的相关物质、特性、反应等，以及生物医药实验的相关技术的专有名词；而模型 B 抽取的实体主要为蛋白质、基因、细胞的相关物质的专有名词，大多数实体的字符特征包含数字、大写字母或希腊字符（如 ε、α、β 等）。

从候选实体的构成词数量来看，模型 G 通常会将该词的上文或下文的词共同抽取出来作为多词实体，而模型 B 通常仅抽取单个词作为实体。

从词在应用集中被标注为实体构成词的概率来看，同一个词在模型 B 和模型 G 中被标注的标签分布概率有较大不同，特别是非基因类的词，如“molecules”（分子）在模型 G 中作为应用集的实体构成词的概率为 75.86%，而在模型 B 中仅为 13.79%。对于基因名实体，在两个模型中的预测结果则相当接近，分别为 100% 和 90%。

经过修正后，部分 S^{ent-e} 低于 0.5 的候选实体被筛除。由此可见，采用基于实体构成词概率进行筛选的方法能够筛掉一些不属于医药实体的候选项。但是当训练集的实体词比例极少时，候选实体的实体构成词的概率大多数为应用集中的概率，若应用集中的标注不够准确，则会影响后续对候选实体的筛选。

以模型 G 生成的标注文件和 XML 格式文件为例，展示如下。

标注文件示例：Epigenome-wide O\association

O\studies O\can O\identify O\environmentally O\mediated O\epigenetic O\changes O\such O\as O\altered O\DNAB\methylation E，O。

XML格式文件示例：Epigenome-wide association studies can identify environmentally mediated epigenetic changes such as altered <entity> DNA methylation </entity>，which may also be influenced by genetic factors.

To investigate possible contributions of DNA methylation to the aetiology of <entity>rheumatoid arthritis</entity> with minimum confounding genetic heterogeneity，we investigated <entity>genome-wide <entity>DNA methylation</entity> </entity> in <entity>disease-discordant monozygotic twin pairs</entity>。

为验证模型识别新医药实体的能力，随机选取部分模型识别出的实体并对实体进行词形还原；将该实体输入医药领域的规范主题词表搜索引擎中进行检索，若该实体在主题词表中无相关词条，且经人工判断确实为医药实体，即表示模型发现了新的医药实体。在本实验中，选择由美国国家医学图书馆出版的医学主题词表（Medical Subject Heading，MeSH）作为检索的主题词表。经检索，如RUNX3（RUNX3基因）、TCR（T细胞抗原受体）、glioma cell（神经胶质瘤细胞）等较常见的医药实体在MeSH词表中已有相关记录，而HNSCC cell（头颈鳞状细胞癌细胞）、DNAH10（DNAH10基因）、qMSP（定量甲基化特异性聚合酶链式反应）等新的医药实体或医药实体缩写词则在MeSH词表中则尚未有记录，说明本文模型可辅助领域专家从科技论文中发现和挖掘新出现的医药实体以及未登录的医药实体缩写词，有助于医药科技论文的知识发现和分析利用。

6 结 语

本文提出的Attention-BiLSTM-CRF模型的医药实体识别效果优于基准模型，且在数据极度不平衡的数据集上更有优势。在模型验证中，分别使用以两个公开数据集为训练集的Attention-BiLSTM-CRF模型对医学论文的摘要进行医药实体抽取、语料标注和实体识别。结果显示，受数据集的特性影响，以GENIA数据集为训练集的模型能获得更多种类的实体，且倾向于标注为多词实体；而以BCII-GM数据集为训练集的模型则主要获得基因实体，且倾向于标注为单词实体。被标注的实体大多符合医药实体的概念，说明模型在实际应用中有一定的实用性。模型生成的应用标注文件和应用实体文件在经过进一步的人工修正后，可作为医药领域关系抽取、本体构建、词典构建等研究的语料。此外，本研究提出的模型的实体识别准确率在后续研究中可进一步提升，并且模型在不同数据集中的表现不够稳定，还需要进一步改进。

参考文献：

[1] 张海楠, 伍大勇, 刘悦, 等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报, 2017, 31(4): 28-35.(Zhang Hainan, Wu Dayong, Liu Yue, et al. Chinese Named Entity Recognition Based on Deep Neural Network[J]. Journal of Chinese Information Processing, 2017, 31(4): 28-35.)

[2] 姚霖, 刘轶, 李鑫鑫, 等. 词边界字向量的中文命名实体识别[J]. 智能系统学报, 2016, 11(1): 37-42. (Yao Lin, Liu Yi, Li Xinxin, et al. Chinese Named Entity Recognition via Word Boundary Based Character Embedding[J]. CAAI Transactions on Intelligent Systems, 2016, 11(1): 37-42.)

[3] Bengio Y, Schwenk H, Senécal J S, et al. Neural Probabilistic Language Models[A]//Holmes D E, Jain L C. Innovations in Machine Learning[M]. 2006: 137-186.

[4] Luong T, Pham H, Manning C D. Effective Approaches to Attention-Based Neural Machine Translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1412-1421.

[5] 张帆, 王敏. 基于深度学习的医疗命名实体识别[J]. 计算技术与自动化, 2017, 36(1): 123-127. (Zhang Fan, Wang Min. Medical Text Entities Recognition Method Base on Deep Learning[J]. Computing Technology and Automation, 2017, 36 (1): 123-127.)

[6] 张聪品, 方滔, 刘昱良. 基于LSTM-CRF命名实体识别技术的研究与应用[J]. 计算机技术与发展, 2019, 29(2): 106-108. (Zhang Congpin, Fang Tao, Liu Yuliang. Research and Application of Named Entity Recognition Based on LSTM-CRF [J]. Computer Technology and Development, 2019, 29(2): 106-108.)

[7] 申站. 基于神经网络的中文电子病历命名实体识别[D]. 北京: 北京邮电大学, 2018.(Shen Zhan. Named Entity Recognition for Chinese Electronic Record with Neural Network[D]. Beijing: Beijing University of Posts and Telecommunications, 2018.)

[8] 薛天竹. 面向医疗领域的中文命名实体识别[D]. 哈尔滨: 哈尔滨工业大学, 2017.(Xue Tianzhu. Research on Chinese Named Entity Recognition in Medical Field[D]. Harbin: Harbin Institute of Technology, 2017.)

[9] dos Santos C N, Zadrozny B. Learning Character-Level Representations for Part-of-Speech Tagging[C]//Proceedings of the 31st International Conference on Machine Learning. 2014: 1818-1826.

[10] LeCun Y, Bottou L, Bengio Y, et al. Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[11] Zhao Z H, Yang Z H, Luo L, et al. Disease Named Entity Recognition from Biomedical Literature Using a Novel Convolutional Neural Network[J]. BMC Medical Genomics, 2017, 10(S5): 73.

[12] Elman J L. Finding Structure in Time[J]. Cognitive Science, 1990, 14(2): 179-211.

[13] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[14] Cho K, van Merrienboer B, Bahdanau D, et al. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches [C]//Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation. 2014: 103-111.

[15] Huang D G, Jin L K, Song D X, et al. Biomedical Named Entity Recognition Based on Recurrent Neural Networks with Different Extended Methods[J]. International Journal of Data Mining and Bioinformatics, 2016, 16(1): 17.

[16] Liu Z J, Yang M, Wang X L, et al. Entity Recognition from Clinical Texts via Recurrent Neural Network[J]. BMC Medical Informatics and Decision Making, 2017, 17(S2): 67.

[17] Sahu S, Anand A. Recurrent Neural Network Models for Disease Name Recognition Using Domain Invariant Features[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 2216-2225.

[18] Gridach M. Character-Level Neural Network for Biomedical Named Entity Recognition[J]. Journal of Biomedical Informatics, 2017, 70: 85-91.

[19] Zeng D H, Sun C J, Lin L, et al. LSTM-CRF for Drug-Named Entity Recognition[J]. Entropy, 2017, 19(6): 283.

[20] Habibi M, Weber L, Neves M, et al. Deep Learning with Word Embeddings Improves Biomedical Named Entity Recognition[J]. Bioinformatics, 2017, 33(14): i37-i48.

[21] Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent Neural Networks with Specialized Word Embeddings for Health-Domain Named-Entity Recognition[J]. Journal of Biomedical Informatics, 2017, 76: 102-109.

[22] Devlin J, Chang M W, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding[OL]. arXiv Preprint, arXiv: 1810.04805.

[23] Souza F, Nogueira R, Lotufo R. Portuguese Named Entity Recognition Using BERT-CRF[OL]. arXiv Preprint, arXiv: 1909.10649.

[24] Alsentzer E, Murphy J, Boag W, et al. Publicly Available Clinical [C]//Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019: 72-78.

[25] Lee J, Yoon W, Kim S, et al. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[26] Lyu C, Chen B, Ren Y F, et al. Long Short-Term Memory RNN for Biomedical Named Entity Recognition[J]. BMC Bioinformatics, 2017, 18(1): 462.

[27] Rei M, Crichton G K O, Pyysalo S. Attending to Characters in Neural Sequence Labeling Models[C]//Proceedings of the 26th International Conference on Computational Linguistics. 2016: 309-318.

[28] Luo L, Yang Z H, Yang P, et al. An Attention-Based BiLSTM-CRF Approach to Document-Level Chemical Named Entity Recognition[J]. Bioinformatics, 2018, 34(8): 1381-1388.

[29] Graves A, Schmidhuber J. Framewise Phoneme Classification with Bidirectional LSTM Networks[C]//Proceedings of the 2005 IEEE International Joint Conference on Neural Networks. 2005: 2047-2052.

[30] Milolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[OL]. arXiv Preprint, arXiv: 1301.3781.

[31] Yuan Z, Liu Y J, Tan C Q, et al. Improving Biomedical Pretrained Language Models with Knowledge[C]//Proceedings of the 20th Workshop on Biomedical Language Processing. 2021.

[32] Yuan Z, Tan C Q, Huang S F, et al. Fusing Heterogeneous Factors with Triaffine Mechanism for Nested Named Entity Recognition [OL]. arXiv Preprint, arXiv: 2110.07480.

作者贡献声明：

赵蕊洁：提出研究思路，设计研究方案；
佟昕瑀：进行实验；
刘小桦：采集、清晰和分析数据，起草论文；
路永和：论文最终版本修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储，E-mail: zhaorj8@mail2.sysu.edu.cn。
[1] 赵蕊洁, 刘小桦. Medical dataset.zip. 医药数据集.

[2] 赵蕊洁, 刘小桦. Code.zip. 论文实验源代码.
[3] 赵蕊洁, 佟昕瑀. 实验结果.docx. 实验结果数据.

收稿日期:2021-12-15
收修改稿日期:2022-05-10

Entity Recognition and Labeling for Medical Literature Based on Neural Network

Zhao Ruijie Tong Xinyu Liu Xiaohua Lu Yonghe
(School of Information Management, Sun Yat-Sen University, Guangzhou 510006, China)

Abstract: [**Objective**] This paper proposes a new entity recognition model, aiming to find new knowledge effectively and improve the utilization of medical papers. [**Methods**] We constructed a pharmaceutical entity recognition model based on Attention-BiLSTM-CRF and examined it on the public datasets of GENIA Term Annotation Task and BioCreative II Gene Mention Tagging. We also used the model to annotate abstracts of biomedical scientific papers. [**Results**] The F1 values of our model on the two data sets were 81.57% and 84.23%, while the accuracy rates were 92.51% and 97.85%. These results are better than those of the benchmark ones. Moreover, our model has more advantages in processing the extremely unbalanced data. [**Limitations**] The volume of data and application of entity labeling experiments are relatively homogeneous. [**Conclusions**] The proposed model improves the effectiveness of entity recognition and mining of new medical knowledge.

Keywords: Biomedical Named Entity Recognition Entity Annotation Neural Network Attention Mechanism