计算机科学
COMPUTER SCIENCE

基于多特征嵌入的中文医学命名实体识别

黄健格, 贾真, 张凡, 李天瑞

引用本文

黄健格, 贾真, 张凡, 李天瑞.基于多特征嵌入的中文医学命名实体识别[J]. 计算机科学, 2023, 50(6): 243-250.

HUANG Jiange, JIA Zhen, ZHANG Fan, LI Tianrui. Chinese Medical Named Entity Recognition Based on Multi-feature Embedding [J]. Computer Science, 2023, 50(6): 243-250.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220400115

基于多特征嵌入的中文医学命名实体识别

黄健格[1]　贾　真[1,2]　张　凡[1,2]　李天瑞[1,2,3]
1 西南交通大学计算机与人工智能学院　成都 611756
2 四川省制造业产业链协同与信息化支撑技术重点实验室　成都 611756
3 综合交通大数据应用技术国家工程实验室　成都 611756
(hjgeuraka@163.com)

摘　要　针对基于字符表示的中文医学命名实体识别模型嵌入信息单一、缺失词边界和结构信息的问题,文中提出了一种融合多特征嵌入的医学命名实体识别模型。首先,将字符映射为固定长度的嵌入表示;其次,引入外部资源构建词汇特征,该特征能够补充字符的潜在词组信息;然后,根据中文的象形文字特点和文本序列特点,分别引入字符结构特征和序列结构特征,使用卷积神经网络对两种结构特征进行编码,得到 radical-level 词嵌入和 sentence-level 词嵌入;最后,将得到的多种特征嵌入进行拼接,输入长短期记忆网络编码,并使用条件随机场输出实体预测结果。将自建中文医疗数据和 CHIP_2020 任务提供的医疗数据作为数据集进行实验,实验结果表明,与基准模型相比,所提模型同时融合了词汇特征和文本结构特征,能够有效识别医学命名实体。

关键词:命名实体识别;中文医学文本;词汇信息;文本结构特征;深度学习
中图法分类号　TP391

Chinese Medical Named Entity Recognition Based on Multi-feature Embedding

HUANG Jiange[1],JIA Zhen[1,2],ZHANG Fan[1,2] and LI Tianrui[1,2,3]
1 School of Computing and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China
2 Manufacturing Industry Chains Collaboration and Information Support Technology Key Laboratory of Sichuan Province,Chengdu 611756,China
3 National Engineering Laboratory of Integrated Transportation Big Data Application Technology,Chengdu 611756,China

Abstract　Aiming at the problems of single embedding information,lacking of word boundary and text structure information in Chinese medical named entity recognition(NER) model based on character representation,this paper presents a medical named entity recognition model integrating multi-feature embedding. Firstly,the characters are mapped to a fixed-length embedding representation. Secondly,external resources are introduced to construct lexical feature,which can supplement the potential phrase information of characters. Thirdly,according to the characteristics of Chinese pictographs and text sequences,character structure feature and sequence structure feature are introduced,respectively. The convolutional neural networks are used to encode the two structural features to obtain radial-level word embedding and sentence-level word embedding. Finally,the obtained multiple feature embeddings are concatenated and input into the long short-term memory network encoding,and the entity result is output by the CRF layer. Taking the self-built Chinese medical data and the CHIP_2020 data as the datasets,experimental results show that compared with the benchmark models,the proposed model integrating both lexical feature and text structure feature can effectively identify named entities in the medical field.

Keywords　Named entity recognition,Chinese medical text,Lexical information,Text structure features,Deep learning

1　引言

随着大数据时代的迅速发展,医学产业逐渐实现现代化管理,同时也产生了大量详细且有价值的医学研究数据,如电子病历、医学文献等。但是,多数医学数据以非结构化的形式存在,计算机不能直接处理这些数据。因此,生物医学研究考虑借助信息抽取技术将电子病历等文本转换为特定的知识,以便相关从业人员能够准确高效地利用有效信息。其中,

到稿日期:2022-04-11　返修日期:2022-09-15
基金项目:国家自然科学基金(62176221)
This work was supported by the National Natural Science Foundation of China(62176221).
通信作者:李天瑞(trli@swjtu.edu.cn)

生物医学命名实体识别(Biomedical Named Entity Recognition,BioNER)是生物信息抽取技术的研究热点。BioNER 指从非结构化文本中识别命名实体,如基因、症状、医疗设备和药物名称等,该研究成果将为构建医疗知识图谱、药物监管和临床决策提供基础支撑。

在英文医学实体识别数据集上,Cho 等[1]提出在嵌入层组合多种特征,利用单词的字母组成,分别采用局部和全局深度神经网络进行编码,设计基于字母表示和单词表示相结合的嵌入层,有效提高了英文医学领域实体识别的性能。但是,中文文本与英文相比,具有如下两个特点。1)英文单词间存在天然的分隔符,而中文句子的字符是紧密排列的形式。因此,中文命名实体识别的做法是利用已有的分词系统进行分词后编码,或者结合分词模型进行联合训练[2-3]。然而,分词的效果很大程度上取决于分词工具,分词工具不可避免地会对词语边界进行错误划分。因此,常用的 BioNER 做法是基于字符粒度建模,但是字符粒度又会缺失潜在的词组语义。2)英文单词由字母组成,而汉字是象形文字,在中文文本中,一般情况下汉字不会像英文单词一样拆解为字母组合。但是,中文字符的结构蕴含着汉字的语义表达,将汉字分解为细粒度的组成,可以提升文本的表示性能。同时,字与字之间的序列结构也存在一定的语义特征。因此,根据上述分析,本文提出了基于多特征嵌入的中文医学命名实体识别模型(Multi-Feature Embedding based on BiLSTM-CRF, MFEBC),在嵌入层设计组合特征,融合字符潜在的词组信息、汉字的字符结构信息和文本的序列结构信息,实现中文文本的有效表示。

2 相关工作

随着医学与信息技术的融合,命名实体识别(Named Entity Recognition,NER)变得越来越重要。其主要的研究方法包括基于规则和词典的方法、基于机器学习的方法和基于深度学习的方法。

最早的命名实体识别主要基于规则和词典[4]来实现。基于规则的方法根据文本数据手工定义规则来识别命名实体。基于词典的方法,将词典中的词汇与目标文本简单匹配,用于实体提取。但实体数量的不断增加给实体提取带来了困难,有限的规则和词典不能匹配领域内所有的实体。

为解决以上问题,机器学习模型逐渐取代了基于规则和词典的方法。基于机器学习的方法主要采用马尔可夫模型[5-6]和条件随机场(Conditional Random Field,CRF)模型[7-9]等。基于机器学习的方法能够识别词典和规则之外的命名实体,但是该方法依赖大量的特征工程,而且需要专业的领域知识。

近年来,深度神经网络被引入到命名实体识别任务中,用于提炼高质量的编码信息,在通用领域取得了显著效果。Dong 等[10]首次将基于字符嵌入的双向长短期记忆网络(Bidirectional Long Short-Term Memory,BiLSTM)和 CRF 相结合并将其用于中文 NER,而且将字符的偏旁视为序列输入 BiLSTM 进行编码,在没有人工设计特征的情况下,取得了较好的效果。Liu 等[11]将汉字的结构视为图像,通过卷积神经网络(Convolutional Neural Network,CNN)编码汉字的图片来提取字符的结构信息。Song 等[12]将自注意力机制和多嵌入技术相结合,从字符、部首等特征中捕获不同粒度的语义表示,找出字符之间的相关性。Zhang 等[13]提出了中文 NER 的 Lattice-LSTM 模型,该模型对词典匹配的潜在词语进行编码,设计门控单元以选择最佳匹配的字符和词语。Ma 等[14]提出了分类匹配词汇信息,该方法具有更快的推理速度和更优的推理性能。Liu 等[15]提出了基于词库增强 BERT 的方法,通过词库适配器将外部词汇直接整合到 BERT 层中。

在生物医学命名实体识别领域,许多研究都采用 BiLSTM-CRF 模型作为核心框架,对生物医学数据进行实体识别。Gridach[16]最先将 BiLSTM-CRF 应用于英文生物医学命名实体识别,并且证明了结合单词和字母的有效性。Yin 等[17]提出了基于偏旁特征和自注意力机制的 BiLSTM-CRF 模型,用于中文医学命名实体识别。Gong 等[18]提出融合字形和字符粒度的特征,来识别电子病历中的实体类型。考虑到中文 BioNER 基于字符嵌入的模型缺失词组的信息,Li 等[19]使用 Lattice-LSTM 提取中文临床医学实体。Zhao 等[20]基于新词发现进行中文医疗实体识别。除此之外,Wang 等[21]提出使用多任务学习框架集中训练多种类型实体。Hu 等[22]提出了基于知识蒸馏的生物医学命名实体识别算法,有效压缩了模型大小,加快了模型推理速度。由于预训练模型在自然语言处理任务上的广泛应用,许多基于 BERT 的拓展研究[23-24]在 BioNER 领域也取得了突破性进展。然而,基于深度学习的方法大多只考虑引入单一的嵌入特征,缺乏深入挖掘文本语义的能力。因此,受 Cho 等[1]的启发,本文提出了一种基于多特征嵌入的中文医学命名实体识别模型,组合词汇特征和文本的结构特征,以达到有效识别命名实体的目的。

3 MFEBC 模型

本文的实体识别模型的总体结构如图 1 所示。MFEBC 模型主要分为嵌入层、编码层和输出层。首先,在嵌入层引入字符信息、词组信息和文本结构信息,分别得到字嵌入、词汇嵌入和文本结构特征嵌入,然后将多种不同层级的嵌入进行拼接,最后输入编码层,由 CRF 输出标记序列的最大概率。同时,本文将预先训练的 BERT 生成词向量作为 BERT 词嵌入。

3.1 嵌入层

3.1.1 词汇嵌入

基于字符嵌入的神经网络模型缺失词组和词边界的信息,字符单独作为嵌入表示会丢失实体词的隐藏关系。因此,在嵌入层引入词汇特征 SoftLexicon[14]来融合词组信息。

本文将命名实体识别作为序列标注任务。序列标注任务可分为“BMES”“BIO”“BIEO”等策略。其中,B 表示实体起始字符,M 或 I 表示实体中间字符,E 表示实体结尾字符,S 表示字符是单个实体,O 表示非实体字符。本文的嵌入层结合 SoftLexicon 来表示,因此采用“BMES”标注的形式,如图 2 所示。

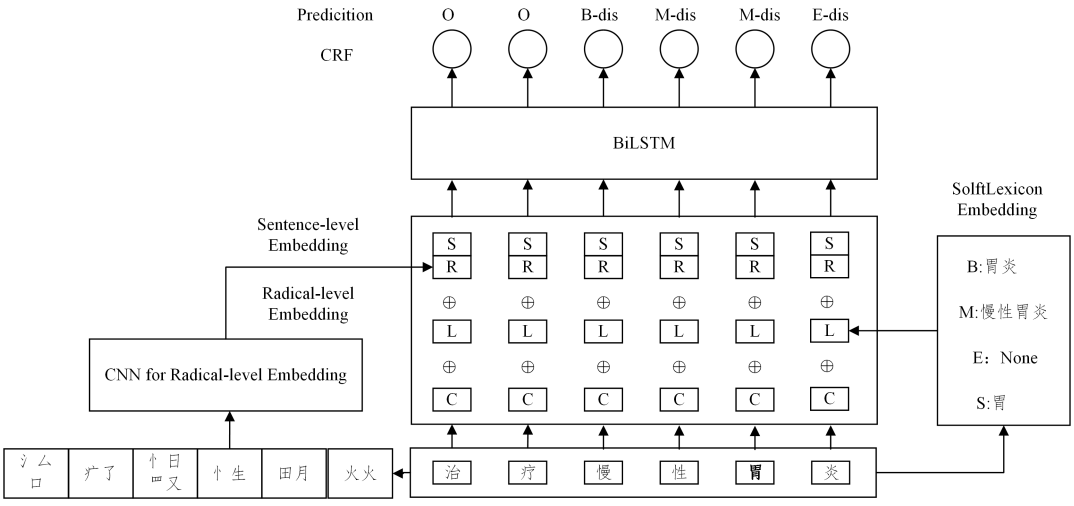

图 1 MFEBC 模型

Fig. 1 MFEBC model

字符序列	牙	龈	炎	患	者	可	服	用	维	生	素	C
标注序列	B-dis	M-dis	E-dis	O	O	O	O	O	B-dru	M-dru	M-dru	E-dru
实体类型	疾病类型								药物类型			

图 2 标注形式举例

Fig. 2 Example of labeling scheme

为了保留词语的边界信息以及词组的语义信息，首先将字符序列 $s=\{c_1,c_2,\cdots,c_{n-1},c_n\}$ 和词典进行匹配，将字符 c_i 匹配得到的词组结果分为“BMES”4 类，构造方法如式(1)所示：

$$\begin{aligned}B(c_i)&=\{w_{i,k},\forall w_{i,k}\in L,i<k\leqslant n\}\\M(c_i)&=\{w_{j,k},\forall w_{j,k}\in L,1\leqslant j<i<k\leqslant n\}\\E(c_i)&=\{w_{j,i},\forall w_{j,i}\in L,1\leqslant j<i\leqslant n\}\\S(c_i)&=\{c_i,\exists c_i\in L,1\leqslant i\leqslant n\}\end{aligned}\tag{1}$$

其中，L 表示词典集合，w 表示句子和词典匹配得到的词语，n 表示文本长度。

如图 3 所示，以“治疗肠胃炎”为例。字符“胃”字和词典匹配的结果集合分别是：B 集合包含 B(胃)=$\{w_{4,5}$(胃炎)$\}$，M 集合包含 M(胃)=$\{w_{3,4}$(肠胃炎)$\}$，E 集合包含 E(胃)=$\{w_{3,4}$(肠胃)$\}$，S 集合包含 S(胃)=$\{c_4$(胃)$\}$。如果某个分类集合没有匹配词语，则用 *None* 表示，如 c_5(炎)是句子的结尾字符，没有以“炎”作为开始的匹配词语，则 B(炎)=$\{None\}$。

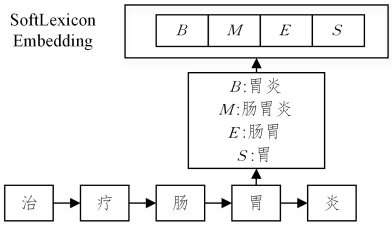

图 3 词典匹配

Fig. 3 Lexicon matching

在得到字符的词典匹配结果之后，将每一类集合映射为固定长度的词嵌入，但是可能存在同一集合内有多个匹配词语的情况，模型使用静态加权统计的方法，将多个词语映射到固定长度的向量。使用静态数据得到词语的词频并将其直接作为权重，静态数据统计包括训练集和验证集的医疗文本数据，加权方法如式(2)所示：

$$\boldsymbol{V}^s(S)=\frac{4}{z}\sum_{w\in T}z(w)\boldsymbol{e}^w(w)\tag{2}$$

其中，T 表示用于权重统计的文本，$\boldsymbol{e}^w(w)$ 表示词向量，$z(w)$ 表示词频，z 表示 4 个词集的词频总数。词典集合使用 Chinese Giga-Word[13] 训练得到的词向量词典。

最后，得到字符 c_i 的“BMES”词组集合，拼接 4 个集合的嵌入表示，得到固定长度的输入向量，如式(3)所示；

$$\boldsymbol{e}^c(B,M,E,S)=[\boldsymbol{v}^c(B)\oplus\boldsymbol{v}^c(M)\oplus\boldsymbol{v}^c(E)\oplus\boldsymbol{v}^c(S)]\tag{3}$$

其中，$\boldsymbol{e}^c(B,M,E,S)$ 表示字符 c_i 和词典匹配后的 SoftLexicon 向量，$\boldsymbol{v}^c$ 表示单个集合内部加权之后的向量，$\oplus$ 表示拼接。

3.1.2 文本结构特征嵌入

文本结构特征嵌入分为字符结构嵌入(Radical-level Embedding)和序列结构嵌入(Sentence-level Embedding)。

在英语中，单词可以拆分为字母，不同的单词可能具有相同的词根和词缀。词根和词缀对单词的含义通常具有决定性意义，例如“peps”和“pept”词根表示“消化”，相关单词有 dyspeptic(消化不良的)、peptic(助消化的)、eupepsia(消化良好)等[25]。同样地，与英文相似，汉字是由象形文字发展而来的，汉字的内部结构在一定程度上具有形态学信息。特别是医疗领域，如“症”由“疒”和“正”组成，本义是病症，“疒”其古文字形体像病床，常与疾病相关，表示病象，“正”有纠正之义，表示病症须医治。而且，相似的字符结构具有相近的含义，如“月”作为偏旁时，意思和人的器官有关，“肺”“肚”“胃”等均表示人体的部位。因此，模型根据中文的象形文字特点得到字符的结构特征。

如表 1 所列，汉字结构可以有多种拆解方法，其中包括首尾分解和部件构造分解。首尾分解将汉字拆为部首和其他部分组成，部件构造分解则将字符拆为更细粒度的偏旁组成。如，“慢”按照首尾分解，由部首“忄”和偏旁“曼”组成；按照

1) http://tool.httpcn.com/Zi/

部件构造分解，则由“忄、日、罒、又”这4个偏旁组成。通常，更加细粒度的汉字拆分方法能够包含更多的结构信息，因此本文选择部件构造分解作为汉字的组成结构。对于数据集的每个字符，通过“在线新华字典”[1]获取字符的部件构造分解，并且使用卷积神经网络提取字符结构特征。

表1 汉字分解

Table 1 Chinese character decomposition

字符	首尾分解	部件构造分解
痛	疒 甬	疒 冂 丶 用
慢	忄 曼	忄 日 罒 又
药	艹 约	艹 纟 勺 丶

字符结构特征提取网络如图4所示，采用CNN对汉字的部件构造编码，输入是字符的部件构造序列 $R=\{r_1,r_2,\cdots,r_{l-1},r_l\}$，利用最大池化提取字符的结构特征，计算过程如式(4)所示：

$$r^c=\mathrm{Max}\ p\left(\tanh\left(Conv\left(\begin{bmatrix}\boldsymbol{e}^r(r_1)\\ \cdots\\ \boldsymbol{e}^r(r_l)\end{bmatrix}\right)\right)\right) \tag{4}$$

$$\tanh(x)=\frac{\boldsymbol{e}^x-\boldsymbol{e}^{-x}}{\boldsymbol{e}^x+\boldsymbol{e}^{-x}}$$

其中，$\boldsymbol{e}^r$表示部件构造的向量查找表，r_*表示字符c_i的部件构造组成，l表示字符c_i的部件构造数目，tanh表示激活函数，Conv表示卷积操作，Max p表示最大池化，$\boldsymbol{r}^c$表示c_i的字符结构特征嵌入。

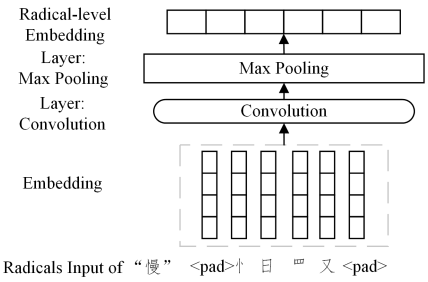

图4 字符结构特征嵌入网络

Fig. 4 Character structure feature embedding network

在序列中，字与字之间的位置结构对字符的表示也存在一定的语义增强作用，如“炎”之前的字符常和“炎”组成疾病名称，即“唇炎”“慢性咽炎”。因此，本文使用CNN对文本进行编码，如图5所示，CNN的输入是字符序列 $s=\{c_1,c_2,\cdots,c_{n-1},c_n\}$，使用卷积层获取字与字之间的序列结构特征。

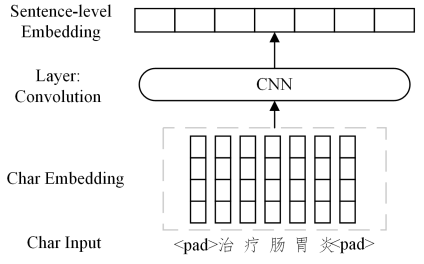

图5 文本结构特征嵌入网络

Fig. 5 Text structure feature embedding network

3.1.3 嵌入层拼接

如式(5)所示，将获取的多种特征嵌入拼接为固定长度的向量，将向量作为整体输入编码层。

$$\boldsymbol{x}^c=[\boldsymbol{x}^c\oplus\boldsymbol{e}^c(B,M,E,S)\oplus\boldsymbol{r}^c\oplus\boldsymbol{s}^c] \tag{5}$$

其中，$\boldsymbol{x}^c$表示字符嵌入，$\boldsymbol{r}^c$表示字符结构特征嵌入，$\boldsymbol{s}^c$表示序列结构特征嵌入。

3.2 编码层

长短期记忆网络(Long Short-Term Memory，LSTM)是循环神经网络(Recurrent Neural Network，RNN)的变体，RNN是一种常用于处理序列数据的神经网络，其考虑序列当前的输出不仅和节点自身有关，也和前面节点的输出有关。LSTM是一种特殊的RNN，主要是为了解决训练过程中梯度爆炸和梯度消失的问题，与RNN相比，LSTM在长序列中有更好的表现。LSTM由多个LSTM单元组成，每个LSTM单元由输入门、遗忘门、输出门以及细胞状态组成。嵌入层的输入向量经过LSTM单元处理得到文本的编码结果，LSTM单元的计算过程如式(6)所示：

$$\begin{aligned}
\boldsymbol{i}_t&=\sigma(\boldsymbol{W}_{x_i}\boldsymbol{X}_t+\boldsymbol{W}_{h_i}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i)\\
\boldsymbol{f}_t&=\sigma(\boldsymbol{W}_{x_f}\boldsymbol{X}_t+\boldsymbol{W}_{h_f}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f)\\
\boldsymbol{o}_t&=\sigma(\boldsymbol{W}_{x_o}\boldsymbol{X}_t+\boldsymbol{W}_{h_o}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o)\\
\boldsymbol{c}_t&=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tanh(\boldsymbol{W}_{x_C}\boldsymbol{X}_C+\boldsymbol{W}_{h_C}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c)\\
\boldsymbol{h}_t&=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_o)
\end{aligned} \tag{6}$$

其中，σ表示sigmoid函数，$\boldsymbol{W}_*$和$\boldsymbol{b}_*$表示权重矩阵和偏置项，$\boldsymbol{i}_*$，$\boldsymbol{f}_*$，$\boldsymbol{o}_*$分别表示输入门、遗忘门和输出门，$\odot$表示按元素点积。

单向LSTM只能捕获前文的序列信息，而缺失后文的信息，因此编码层采用BiLSTM。BiLSTM包含前向和后向两个传播过程，可以捕获序列的上下文信息。在t时刻，获得前向隐藏状态$\overrightarrow{\boldsymbol{h}_t}$和后向隐藏状态$\overleftarrow{\boldsymbol{h}_t}$，连接两种状态得到$t$时刻的隐藏状态$\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}_t}:\overleftarrow{\boldsymbol{h}_t}]$。

3.3 CRF

虽然BiLSTM得到的概率矩阵经过Softmax输出可以判断最终预测结果，但是命名实体识别的序列标签存在依赖关联(如B-pro后面不能跟M-sym等)，使用CRF可以利用序列标注的前后联系，在全局意义上给标签添加约束，得到最大概率的合理序列。将BiLSTM编码结果 $X=\{x_1,x_2,\cdots,x_{n-1},x_n\}$作为CRF的输入，设CRF的输出序列为 $y=\{y_1,y_2,\cdots,y_{n-1},y_n\}$，那么得分公式如式(7)所示：

$$score(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{7}$$

其中，$\boldsymbol{A}$表示转移矩阵，元素$A_{y_i,y_{i+1}}$表示标签y_i到y_{i+1}的概率得分，P_{i,y_i}表示第i个序列被标注为y_i的概率。最后在标签的概率得分上使用Softmax得到归一化的条件概率，如式(8)所示：

$$P(y|\boldsymbol{X})=\frac{\mathrm{e}^{score(\boldsymbol{X},y)}}{\sum\limits_{\tilde{y}\in y}score(\boldsymbol{X},\tilde{y})} \tag{8}$$

其中，y表示所有可能输出的标签，$\tilde{y}$表示真实标签。在训练中，采用最大似然估计作为损失函数，得到最大对数概率，如式(9)所示：

$$\log P(y|\boldsymbol{X})=score(\boldsymbol{X},y)-\log\Big(\sum_{\tilde{y}\in\boldsymbol{Y}_x}\mathrm{e}^{score(\boldsymbol{X},\tilde{y})}\Big) \tag{9}$$

在解码阶段，使用动态规划，搜寻最大的条件概率分数，通过式(10)计算结果。

$$y^* = \text{argmax}\ score(\boldsymbol{X}, \tilde{y}) \tag{10}$$

4 实验

4.1 实验数据

实验使用两个中文医学实体识别数据集。数据集 I 是自建的医学实体识别数据集，主要来自药品说明书、权威数据和相关文献等，在未处理的文本上进行标注，得到药物、药物类别、病症、建议、给药途径、功能主治、用药剂量、用药疗程这 8 类实体，总计 7000 条数据。数据集 II 来自 Biendata CHIP_2020[1)] 命名实体识别任务，包含疾病、临床表现、医疗程序、医疗设备、药物、检查项目、身体部位、科室、微生物这 9 类实体，将原始数据转换为“BMES”序列标注的形式，得到约 20 000 条数据。将数据集 I 和数据集 II 分别按照 7∶1.5∶1.5 的比例划分为训练集、验证集和测试集，统计结果如表 2 所列。

表 2 数据集的统计结果

Table 2 Statistical results of datasets

数据集	类型	训练集	验证集	测试集
数据集 I	句子	5180	1110	1110
	实体	10010	2220	2130
数据集 II	句子	14000	3000	3000
	实体	53200	11400	11200

4.2 评价指标

模型评估采用准确率(P)、召回率(R)和 F1 值作为评价指标，如式(11)所示：

$$\begin{aligned} P &= \frac{\text{预测正确的实体个数}}{\text{预测实体的总个数}} * 100\% \\ R &= \frac{\text{预测正确的实体个数}}{\text{标注实体的总个数}} * 100\% \\ F1 &= \frac{2 * P * R}{P + R} * 100\% \end{aligned} \tag{11}$$

4.3 实验设置

本文将字符嵌入维度、词组嵌入维度、字符结构嵌入维度设置为 50，卷积核大小设为 3，CNN 输出通道分别设为 150 和 50，BiLSTM 隐藏层大小设为 300，batch-size 设为 16，dropout 设为 0.5，学习率设为 0.0015，使用 Adam 优化器进行训练。

4.4 实验结果

4.4.1 基准模型

为了验证本文模型的性能，本文实现了多种医学命名实体识别模型，并将其作为基准模型。

(1)BiLSTM-CRF：基于字符嵌入，利用 BiLSTM 对上下文进行编码，联合 CRF 对模型标签进行修正。在此基础上，本文将多种嵌入表示的模型进行对比。其中，char＋bichar[26] 添加 bigrams 向量表示作为字符嵌入的补充；Char＋charCNN[27] 将字符嵌入和经过 CNN 处理的编码结果拼接送入 BiLSTM；Char＋charCNN＋charLSTM 与 Cho 等[1] 使用的 charCNN 和 charLSTM 结构类似，本文实现的模型输入序列为中文字符。

(2)BiLSTM＋GCN＋CRF：利用分词后的邻接矩阵构建图神经网络。

(3)BiLSTM-att-CRF：联合 BiLSTM-CRF 和自注意力机制，以获取句子的长距离依赖。

(4)TENER[28]：使用结合相对位置信息的编码器，改进 Transformer 适用于 NER 任务。

(5)LR-CNN[29]：使用卷积神经网络重新合并词汇，对所有与句子匹配的潜在词语进行建模，设计特征优化网络，解决词组冲突。

(6)SoftLexicon[14]：与词典匹配，引入词汇特征。

(7)BERT＋BiLSTM-CRF：添加已经训练好的 BERT 生成词向量，并将其作为 BERT 嵌入。

4.4.2 对比实验

本文在两个数据集上进行实验，实验结果表明，相比基线模型，本文方法的 F1 值有不同程度的提升。

数据集 I 的实验结果如表 3 所列。首先，与 BiLSTM-CRF 相比，本文的 MFEBC 模型的准确率、召回率和 F1 值分别提升了 1.71%，4.47% 和 3.07%，多特征引入更丰富的嵌入表示，模型的整体性能得到提升。TENER 没有引入外部信息，而是使用 Transformer 结合方向和距离编码相对位置，相比 TENER，MFEBC 的 F1 值提升了 2.07%。与 SoftLexicon 和 LR-CNN 模型相比，MFEBC 模型的 F1 值分别提升了 1.31% 和 0.75%，MFEBC 在引入词组信息的基础上，添加了两种文本结构特征，在一定程度上增强了文本的深层语义表示。其次，通过引入 charCNN 和 charLSTM 的嵌入特征向量，模型的效果相对 BiLSTM-CRF 都有不同程度的提升，这说明通过在嵌入层引入多种额外特征能有效提升模型的性能。与 BiLSTM-CRF 相比，BERT＋BiLSTM-CRF 能更好地利用上下文信息，进行动态建模，语义更加精确，F1 值提升了 5.33%。相比 BERT＋BiLSTM-CRF 模型，将 MFEBC 和 BERT 相结合后的 F1 值提升了 0.28%。

表 3 数据集 I 的实验结果

Table 3 Experimental results in dataset I

(单位：%)

Model	*P*	*R*	*F*1
BiLSTM-CRF	85.73	85.85	85.79
＋bichar	81.84	90.04	85.74
＋charCNN	85.63	86.84	86.23
＋charCNN＋charLSTM	86.09	86.94	86.51
BiLSTM-GCN-CRF	84.39	86.41	85.39
BiLSTM-att-CRF	84.52	84.68	84.60
TENER	83.65	90.18	86.79
LR-CNN	86.06	90.27	88.11
SoftLexicon	86.68	88.43	87.55
MFEBC	**87.44**	**90.32**	**88.86**
BERT＋BiLSTM-CRF	90.15	**92.10**	91.12
MFEBC＋BERT	**90.74**	92.06	**91.40**

数据集 II 的实验结果如表 4 所列。MFEBC 模型的 F1 值为 63.08%，与 BiLSTM-CRF 相比，MFEBC 模型的准确率、召回率和 F1 值分别提升了 1.57%，5.21% 和 3.37%。与 TENER，SoftLexicon，LR-CNN 相比，MFEBC 的 F1 值分别

1) www.biendata.xyz

提升了2.34%，1.09%和1.51%，这说明本文引入的多特征信息在数据集II上同样有效。在引入BERT预训练模型后，实验结果表明MFEBC+BERT模型的F1值为65.98%，相比BERT+BiLSTM-CRF提升了0.76%，这说明本文引入的多特征信息可以和BERT预训练模型有效结合。

表4 数据集II的实验结果

Table 4 Experimental results in dataset II

（单位：%）

模型	P	R	F1
BiLSTM-CRF	60.43	59.00	59.71
+bichar	58.77	62.73	60.69
+charCNN	60.89	59.78	60.33
+charCNN+charLSTM	60.91	60.42	60.66
BiLSTM-att-CRF	60.61	60.89	60.75
TENER	59.68	61.84	60.74
LR-CNN	60.59	62.58	61.57
SoftLexicon	**62.05**	61.94	61.99
MFEBC	62.00	**64.21**	**63.08**
BERT+BiLSTM-CRF	64.98	65.46	65.22
MFEBC+BERT	**65.69**	**66.28**	**65.98**

4.4.3 细粒度实体实验结果

为了验证模型在每个实体类别中的实验效果，本文利用MFEBC和BiLSTM-CRF模型在两个数据集的细粒度实体类型上进行实验，评价指标为准确率(P)、召回率(R)和F1值。

模型在数据集I上的细粒度实验结果如表5所列。两种模型在“药物”和“功能主治”类实体上的表现较好。其中，在“药物”类实体的F1值分别达到了92.25%和94.27%。对比两种模型，MFEBC所有实体类型的F1值均有所提升，“用药剂量”和“病症”类实体的提升效果最大，F1值分别提升了4.26%和4.83%。经过分析发现，“用药剂量”类实体多由数字和字符组成，如实体“0.1～0.2克每日”在BiLSTM-CRF模型中会被预测为非实体类(O)，但是在引入多特征信息后会被正确识别为“用药剂量”类。“病症”类型的实体内部包含较多“疒”部首的字符，如“病”“痛”“疡”和“疾”等，F1值的提高，说明引入字符结构特征能够帮助模型正确识别该类实体。

数据集II的细粒度实验结果如表6所列。对比两种模型，MFEBC在所有实体类型上的F1值同样有所提升，其中“医疗设备(equ)”和“部门(dep)”提升最大，F1值分别提升了11.7%和8%。两种模型在“药物(dru)”“疾病(dis)”和“微生物(mic)”类实体上的表现相对较好，其中“药物(dru)”类实体F1值分别达到了76.17%和74.71%。但是对比两个数据集的实验结果，数据集II的实验结果明显偏低。经过分析发现，数据集I的实体组成比较规律，组合为通常意义的名词形式，而数据集II实体组成较为复杂，由较多长难句组成。比如，“机体的生化反应和代谢出现异常”被标注为“临床表现(sym)”，“休息时保持头部抬高30°的卧床位置”被标注为“医疗程序(pro)”，“聚合酶链反应(PCR)技术”被标记为“医学检查项目(ite)”。此外，数据集II中存在实体嵌套的现象，“身体部位(bod)”实体常嵌套在“临床表现(sym)”实体中，如“左心室流出道梗阻”嵌套“左心室”，导致实体F1值较低。“医疗设备(equ)”存在较多英文缩写的情况，导致对语义理解产生障碍。而“科室(dep)”类型的实体受限于部门划分，数据量较少，仅占数据集的0.3%，导致不能充分学习该类实体特征。

表5 数据集I的细粒度实体识别实验结果

Table 5 Experimental results of fine-grained entity recognition in dataset I

（单位：%）

细粒度实体	BiLSTM-CRF			MFEBC		
	P	R	F1	P	R	F1
用药剂量	83.62	81.51	82.55	**87.93**	**85.71**	**86.81**
药物类别	**79.78**	68.93	73.96	74.07	**77.67**	**75.83**
药物	92.49	92.01	92.25	**94.21**	**94.33**	**94.27**
病症	77.68	84.36	80.88	**81.67**	**90.18**	**85.71**
功能主治	**92.26**	79.81	85.58	91.92	**87.50**	**89.66**
给药途径	**86.54**	77.59	81.82	82.76	**82.76**	**82.76**
建议	**89.62**	86.33	87.94	85.77	**90.32**	**87.99**
用药疗程	**91.67**	75.00	82.50	82.35	**87.50**	**84.85**

表6 数据集II的细粒度实验结果

Table 6 Experimental results of fine-grained entity recognition in dataset II

（单位：%）

细粒度实体	BiLSTM-CRF			MFEBC		
	P	R	F1	P	R	F1
dis	67.88	72.43	70.08	**71.35**	**76.83**	**73.99**
sym	**58.52**	43.22	49.72	56.28	**47.62**	**51.59**
pro	56.13	54.24	55.17	**57.58**	**60.80**	**59.15**
equ	43.28	21.97	29.15	**46.60**	**36.36**	**40.85**
dru	74.10	75.33	74.71	**74.88**	**77.50**	**76.17**
ite	**44.80**	28.39	34.76	39.64	**36.45**	**37.98**
bod	56.50	62.31	59.26	**59.04**	**67.36**	**62.93**
dep	60.71	48.57	53.97	**61.11**	**62.86**	**61.97**
mic	69.58	71.95	70.75	**69.82**	**82.76**	**75.74**

4.4.4 实体长度影响

本文对比了实体长度对MFEBC和BiLSTM-CRF模型的影响。实验结果如表7所列，对于不同长度的实体，本文模型的性能均有不同幅度的提升。在数据集I上，当数据长度大于3时，MFEBC模型提升较大，分别提升了6.29%，4.1%，7.4%和8.71%。在数据集II上，长度为5的实体提升最大，F1值提升了5.25%，其余实体也有不同程度的提高，说明引入多特征信息能够有效提升模型性能。

表7 实体长度的实验结果F1值

Table 7 Entity length experiment results F1 value

（单位：%）

dataset	Model	Entity Length						大于等于7
		1	2	3	4	5	6	
数据集I	BiLSTM-CRF	90.80	88.50	89.32	75.70	79.31	71.43	65.09
	MFEBC	**92.73**	**91.36**	**90.20**	**81.99**	**83.41**	**78.83**	**73.80**
数据集II	BiLSTM-CRF	64.02	64.14	60.91	59.87	55.67	57.90	44.63
	MFEBC	**67.49**	**66.95**	**64.80**	**62.80**	**60.92**	**61.86**	**46.75**

4.4.5 Radical-level embedding 可视化

本文在数据集 I 上得到 CNN 训练后的字符结构特征嵌入，将嵌入向量采用主成分分析进行降维，得到二维向量并且可视化。如图 6 所示，“肺”“胃”和“肝”等都有部首“月”，具有相同部首或者结构的汉字在二维空间上距离更近，说明字符结构在一定程度上能增强语义，提升模型性能。

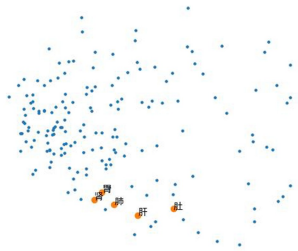

图 6 字符结构嵌入可视化

Fig. 6 Visualization of radical-level embedding

4.4.6 消融实验

为了验证各个模块的有效性，本文在数据集 I 上进行了消融实验。实验结果如表 8 所列，去除 sentence-level embedding 模块后模型的 F1 值下降了 0.39%，去除 SoftLexicon embedding 模块后，模型的 F1 值下降了 1.38%，去除 radical-level embedding 模块后模型性能下降了 1.1%。从实验结果可以看出，去除各个模块后模型的 F1 值有不同程度的下降，验证了本文模型各个模块的合理性和有效性。

表 8 消融实验结果

Table 8 Ablation experiment results

（单位：%）

model	*R*	*P*	*F*1
MFEBC	**87.44**	90.32	**88.86**
-sentence-level embedding	86.39	**90.65**	88.47
-softLexicon embedding	86.18	88.82	87.48
- radical-level embedding	86.95	88.58	87.76

结束语 本文针对中文医学文本特点和字符粒度嵌入信息单一的问题，在嵌入层融合词汇特征和文本结构特征，将多种特征向量拼接输入 BiLSTM-CRF 进行编码，最终模型在两个数据集上的 F1 值均有不同程度的提升，证明了本文模型能够有效识别文本中的医疗实体。但是，医学领域存在较多的嵌套实体，现有的词典还不够完善，导致词典只能匹配实体内部较短的词组。在后续工作中，考虑研究制作大规模中文医学专用词典，补充更加丰富的医学词汇，以提高嵌套实体的识别效果。

参 考 文 献

[1] CHO M，HA J，PARK C，et al. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition[J]. Journal of Biomedical Informatics，2020，103(1)：1-8.

[2] WU F Z，LIU J X，WU C H，et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation [C]// Proceedings of the World Wide Web Conference. 2019：3342-3348.

[3] YANG J，TENG Z Y，ZHANG M S，et al. Combining discrete and neural features for sequence labeling [C] // International Conference on Intelligent Text Processing and Computational Linguistics. Cham，Switzerland：Springer，2016：140-154.

[4] CUI B W，JIN T，WANG J M. Overview of information extraction of free-text electronic medical records[J]. Journal of Computer Applications，2021，41(4)：1055-1063.

[5] AZERAF E，MONFRINI E，VIGNON E，et al. Highly fast text segmentation with pairwise markov chains[C]// Proceedings of the 6th IEEE Congress on Information Science and Technology (CIST). NEW YORK：IEEE，2021：361-366.

[6] HARSHITHA C P，SUNITHAR N R. Topic identification for semantic grouping based on hidden markov model[C]// Proceedings of the 5th International Conference on Communication and Electronics Systems (ICCES). NEW YORK：IEEE，2020：932-937.

[7] SONG S L，ZHANG N，HUANG H T. Named entity recognition based on conditional random fields[J]. Cluster Computing，2019，22(3)：5195-5206.

[8] GONG L J. ZHANG Z F. Clinical named entity recognition from Chinese electronic medical records using a double-layer annotation model combining a domain dictionary with CRF[J]. Chinese Journal of Engineering. 2020，42(4)：469-475.

[9] LIU S，HE T，DAI J. A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese[J]. Mobile Networks and Applications，2021，26(5)：1891-1903.

[10] DONG C H，ZHANG J J，ZONG C Q，et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition [M]// Natural Language Understanding and Intelligent Applications. Cham：Springer，2016：239-250.

[11] LIU F，LU H，LO C，et al. Learning character-level compositionality with visual features[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics，ACL 2017. Vancouver，2017：2059-2068.

[12] SONG C J，XIONG Y，HUANG W C，et al. Joint self-attention and multi-embeddings for Chinese named entity recognition [C]// Proceedings of the 6th International Conference on Big Data Computing and Communications(BIGCOM). New York：IEEE Press，2020：76-80.

[13] ZHANG Y，YANG J. Chinese NER using Lattice LSTM [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1：Long Papers). Stroudsburg：ACL Press，2018：1554-1564.

[14] MA R T，PENG M N，ZHANG Q，et al. Simplify the usage of lexicon in Chinese NER [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL Press，2020：5951-5960.

[15] LIU W，FU X Y，ZHANG Y，et al. Lexicon Enhanced Chinese Sequence Labelling Using BERT Adapter[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(Volume 1：Long Papers). Online：Association for Computational Linguistics，2021：5847-5858.

[16] GRIDACH M. Character-level neural network for biomedical named entity recognition[J]. Journal of Biomedical Informatics，2017，70(5)：85-91.

[17] YINMW, MOUCJ, XIONGKN, etal. Chineseclinicalnamedentityrecognition with radical-level feature and self-attention mechanism [J]. Journal of Biomedical Informatics, 2019, 98(9): 1-7.

[18] GONG D W, ZHANG Y K, GUO Y N, et al. Named entity recognition of Chinese electronic medical records based on multifeatured embedding and attention mechanism [J]. Chinese Journal of Engineering, 2021, 43(9): 1190-1196.

[19] LI Y B, WANG X H, HUI L H, et al. Chinese Clinical Named Entity Recognition in Electronic Medical Records: Development of a Lattice Long Short-Term Memory Model with Contextualized Character Representations[J]. JMIR Medical Informatics, 2020, 8(9): 1-16.

[20] ZHAO Y Q, CHE C ZHANG Q. Chinese medical named entity recognition based on new word discovery and Lattice-LSTM[J]. Computer Applications and Software. 2021(1): 161-165.

[21] WANG X, ZHANG Y, REN X, et al. Cross-type biomedical named entity recognition with deep multi-task learning [J]. Bioinformatics, 2019, 35(10): 1745-1752.

[22] HU B, GENG T Y, DENG G, et al. Faster biomedical named entity recognition based on knowledge distillation[J]. Journal of Tsinghua University (Science and Technology), 2021, 61(9): 936-942.

[23] PENG Y F, YANG S K, LU Z Y. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets[C]// Proceedings of the 18th BioNLP Workshop and Shared Task. Florence: ACL, 2019: 58-65.

[24] GU Y, TINN R, CHENG H, et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM Transactions on Computing for Healthcare (HEALTH), 2021, 3(1): 1-23.

[25] WU S, SONG X N, FENG Z H. MECT: multi-metadata embedding based cross-transformer for Chinese named entity recognition[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2021: 1529-1539.

[26] YANG J, ZHANG Y, DONG F. Neural word segmentation with rich pretraining[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: ACL, 2017: 839-849.

[27] MA X Z, HOVY E. End-to-end Sequence labeling via Bi-directional LSTM-CNNs-CRF[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL Press, 2016: 1064-1074.

[28] YAN H, DENG B, LI X, et al. TENER: adapting transformer encoder for named entity recognition[J]. arXiv: 1911. 04474, 2019.

[29] GUI T, MA R, ZHANG Q, et al. CNN-Based Chinese NER with Lexicon Rethinking[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2019: 4982-4988.

HUANG Jiange, born in 1996, postgraduate, is a member of China Computer Federation. His main research interests include named entity recognition and natural language processing.

LI Tianrui, born in 1969, Ph.D, professor, Ph.D supervisor, is a distinguished member of China Computer Federation. His main research interests include big data intelligence, rough sets and granular computing.

(责任编辑:喻藜)