DOI:10.20033/j.1003-7241.(2025)08-0071-05

BiMcGRU在医疗病历命名实体关系识别中的应用

胡志坚

（福建医科大学附属协和医院，福建 福州 350001）

摘　要：为提升医疗病历实体命名的效果，对基于词汇增强的平格变压器（flat lattice transformer，FLAT)模型进行改进，并利用基于双向神经网络改进的共享多向单元（bi-directional multi-cell GRU，BiMcGRU）模型对传统命名实体识别系统进行优化。在脑血管疾病数据集中，相比于传统FLAT模型，汉字部首特征引入后FLAT模型精确率、召回率和F_1值分别提升了0.91%、0.73%和0.82%。两种实验数据集测试中，基于多任务学习的医疗病历命名实体模型的F_1值分别为89.34%和91.53%，比多任务BERT-BiGRU-ATT-CRF模型的F_1值高，说明BiMcGRU模型能够提升多任务训练识别的效果，研究结果为医疗病历实体命名识别提供新的方法。

关键词：命名实体识别；医疗病历；平格变压器；卷积神经网络；条件随机场

中图分类号：TP183　　**文献标识码：**A　　**文章编号：**1003-7241（2025）08-0071-05

Application of BiMcGRU in Medical Record Naming Entity Relationship Recognition

HU Zhijian

(Fujian Medical University Union Hospital, Fuzhou 350001, China)

Abstract: To improve the effectiveness of medical record entity naming, a study is conducted to improve the flat lattice transformer (FLAT) model based on vocabulary augmentation, and to improve the traditional named entity recognition system using the bi directional Multi cell GRU (BiMcGRU) model based on bidirectional neural network improvement. In the dataset of cerebrovascular diseases, compared to traditional FLAT models, the introduction of Chinese radical features improves the accuracy, recall, and F1 value of the FLAT model by 0.91%, 0.73%, and 0.82%, respectively. In the two experimental datasets tested, the F1 values of the medical record naming entity model based on multi task learning are 89.34% and 91.53%, respectively, which are higher than the F1 values of the multi task BERT BiGRU ATT-CRF model. This indicates that the BiMcGRU model can improve the effectiveness of multi task training and recognition, providing a new method for medical record naming recognition.

Keywords: named-entity recognition; medical records; flat grid transformer; convolutional neural network; conditional random field

0　引言

临床治疗通常使用电子病历来描述患者的医疗情况，并记录患者的疾病及其症状、检查和治疗结果等，电子病历是重要的临床依据[1]。相比于传统纸质病历，电子病历可自动抽取能够高效精确的数据，为医疗诊断提供信息支持，广泛应用于精准医疗研究和疾病监控等[2]。因此，电子病历中有效信息的提取成为研究的重要方向[3-4]。国外对电子病历命名实体识别的研究起步较早，医疗自然语言处理挑战赛推动了该技术的发展。早期电子病历命名实体识别使用基于规则和词典的方法，存在可移植性差和无法解决未登录词汇等问题。与此同时，随着深度学习技术的发展，基于深度学习的方法成为研究热点，如卷积神经网络和循环神经网络模型。例如，预训练语言模型的出现解决了一词多义问题，并提升了实体识别效果。然而，医疗标记语料匮乏且标注难度大，研究人员需要借助多任务学习、迁移学习和主动学习等技术来提高识别精度。国内电子病历命名识别的研究较晚，与国外的任务差别较大。中文电子病历实体命名识别的研究仍存在一些限制，如医疗领域专业词汇较多，现有模型效果不佳；已标注电子病历语料稀缺，无法充分学习文本的语义特征；不同标注医疗语料存在差异，无法合并多个数据集获得大规模医疗数据集；大部分模型依赖人工标注的电子病历语料，忽略了未标注医疗文本中的语义信息[5]。因此，研究在传统命名实体识别基础上，利用深度学习技术进行改进，提出基于词汇增强的平格变压器模型，以及基于双向神经网络改进的共享多向单元模型，期望提升医疗病历命名实体关系识别的效果。

1　医疗病历命名实体关系识别模型设计

1.1　基于部首特征和词汇增强的医疗病历命名实体模型

传统的命名实体识别模型对病历处理的效果并不理

*基金项目：福建省心血管病医学中心建设项目（2021-76）

收稿日期：2023-11-21

想，无法充分利用医疗病历中存在的大量专业名词和复杂句子[6-7]。因此，研究提出一种融合多模块的命名实体识别模型，包含基于词汇增强的平格变压器FLAT模型，卷积神经网络提取部首特征模块、条件随机场（conditional random fields，CRF）层、变压器以及交叉变压器模型，FLAT模型结构如图1所示。

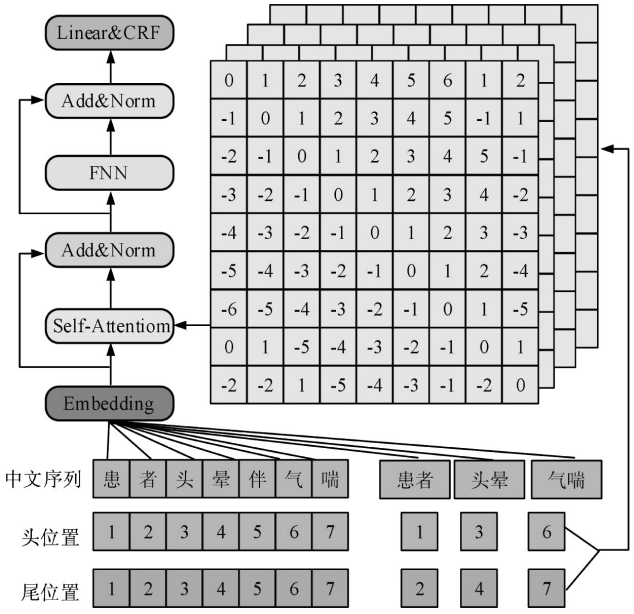


图1 FLAT模型结构

FLAT模型中，Embeding表示植入，Self-Attention表示自适应注意力机制，Add表示残差网络，Norm表示归一化操作，FNN表示前馈网络，Linear表示线性变换，CRF表示条件随机场。FLAT模型序列标注的第1步将词典匹配获得的词汇信息拼接在文本字符后面，并用向量形式表示中文字符和词汇；第2步是通过4种相对位置编码方式得到融合后的字符和词语位置信息。根据头位置和尾位置可计算每两个结点间的四个相对距离，结点的相对位置矩阵如式（1）所示。

$$\begin{cases} \boldsymbol{d}_{ij}^{(hh)} = \text{head}[i] - \text{head}[j] \\ \boldsymbol{d}_{ij}^{(ht)} = \text{head}[i] - \text{tail}[j] \\ \boldsymbol{d}_{ij}^{(th)} = \text{tail}[i] - \text{head}[j] \\ \boldsymbol{d}_{ij}^{(tt)} = \text{tail}[i] - \text{tail}[j] \end{cases} \tag{1}$$

式中，节点x_i的头位置和尾位置分别是head[i]和tail[i]，节点x_i和x_j的头头位置、头尾位置、尾头位置和尾尾位置分别是$\boldsymbol{d}_{ij}^{(hh)}$、$\boldsymbol{d}_{ij}^{(ht)}$、$\boldsymbol{d}_{ij}^{(th)}$和$\boldsymbol{d}_{ij}^{(tt)}$。节点之间的相对位置编码如式（2）所示。

$$\boldsymbol{R}_{ij} = \text{Re lu}(W_r(\boldsymbol{P}_{d_{ij}(hh)} \oplus \boldsymbol{P}_{d_{ij}(ht)} \oplus \boldsymbol{P}_{d_{ij}(th)} \oplus \boldsymbol{P}_{d_{ij}(tt)})) \tag{2}$$

式中，相对位置编码向量为$\boldsymbol{R}_{ij}$，激活函数为Relu，可训练参数为W_r，级联操作为⊕，位置向量为$\boldsymbol{P}_d$。第3步是利用注意力机制获得上下文语义信息，结合CRF模型实现最终序列标注结果的预测。CRF模型生成全局最优标注序列结果如式（3）所示。

$$y_{\text{result}} = \arg\max(S_{(X,Y)}) \tag{3}$$

式中，全局最优标注序列结果为y_{result}，标签计算标签序列得分计算函数为S，输入序列为X，标注结果为Y。研究通过词典匹配的方法，将医疗词汇信息引入FLAT模型，通过位置编码实现字符和词汇信息的融合，以并行化计算方式提高模型训练的效率。

1.2 基于多任务学习的医疗病历命名实体模型

人工构建大规模医疗电子病历标注语料，需要消耗大量的人力和物力进行特征提取和标注[8-10]。因此，为了提升病历标注和特征提取的效率，研究提出一种基于双向神经网络改进的共享多向单元BiMcGRU模型进行共享特征提取，以此构建基于多任务学习的医疗病历命名实体模型，如图2所示。

研究利用BiMcGRU模型对基于变压器的双向编码模型（bidirectional encoder representation from transformers，BERT）内部结构的改进，以提升多任务特征共享的效果。基于变压器的BERT用来完成中文字符的向量化表示，并将向量化表示的中文字符作为BiMcGRU模型的输入。BiMcGRU模型的输出作为各自注意力层的输入，对文本语义信息进行提取，随后通过CRF层进行解码，得到最后的序列标注结果。

模型训练时，多单元门控循环单元（multi-cell gate recurrent unit，McGRU）用来区分实体类型，并通过数据集的实体共享实现知识跨域迁移[11]。McGRU中的组成单元对每个实体类别单元的加权进行计算，实体类型概率

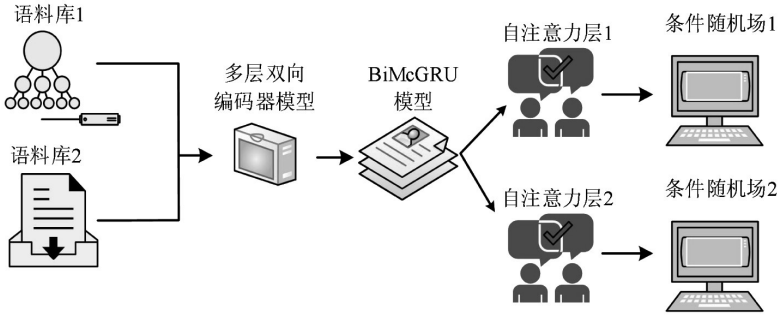


图2 基于多任务学习的医疗病历命名实体模型

决定隐藏层输出单元权重。McGRU实体类型单元中，更新门、重置门和候选隐藏状态的计算如式(4)所示。

$$\begin{cases}\hat{z}_k^{(t)}=\sigma(W_{zk}\cdot[\hat{h}_t^{(t)}\cdot x^{(t)},x^{(t)}]+\boldsymbol{b}_{zk})\\ \hat{r}_k^{(t)}=\sigma(W_{rk}\cdot[\hat{h}_t^{(t-1)}\cdot x^{(t)},x^{(t)}]+\boldsymbol{b}_{rk})\\ \hat{h}_k^{(t)}=\tanh(W_k\cdot[r_t^{(t)}\cdot\hat{h}^{(t-1)},x^{(t)}]+\boldsymbol{b}_k)\end{cases}\tag{4}$$

式中，更新门、重置门和候选隐藏状态分别为$\hat{z}_k^{(t)}$、$\hat{r}_k^{(t)}$和$\hat{h}_k^{(t)}$，门控单元权重分别为W_{zk}、W_{rk}和W_k，偏置矩阵分别为$\boldsymbol{b}_{zk}$、$\boldsymbol{b}_{rk}$和$\boldsymbol{b}_k$。神经网络的隐藏状态如式(5)所示。

$$\hat{h}_\alpha^{(t)}=\sum_{k=1}^{L}(\alpha_k^{(t)}\cdot h_k^{(t)});\text{s.t.}\sum_{k=1}^{L}\alpha_k^{(t)}=1\tag{5}$$

式中，神经网络的隐藏状态为$\hat{h}_\alpha^{(t)}$，第k个隐藏值为$h_k^{(t)}$，对应的权重为$\alpha_k^{(t)}$。真正隐藏状态的计算如式(6)所示。

$$\hat{h}^{(t)}=(1-\hat{z}^{(t)})\cdot\hat{h}^{(t-1)}+\hat{z}^{(t)}\cdot\hat{h}_\alpha^{(t)}\tag{6}$$

式中，真正隐藏状态为$\hat{h}^{(t)}$。

McGRU模型的缺点是不能实现逆向语义信息的获取，只能获取正向历史语音信息[12]。因此，研究借助BiMcGRU模型对正向和逆向记忆单元位置结果进行拼接，并利用上下文信息完成文本特征提取。在BiMcGRU模型训练过程中，可能出现梯度消失的情况，即序列长度的增加会导致模型获取长距离信息能力的下降，自注意力机制的引进便可解决此类问题。

2 医疗病历命名实体关系识别模型实验分析

2.1 基于部首特征和词汇增强的医疗病历命名实体模型实验分析

实验使用的数据集包含全国知识图谱与语义计算大会提供的医疗病历标注数据集、自建脑血管疾病数据库。全国知识图谱与语义计算大会提供的医疗病历标注数据集中包含400份电子医疗病历，总共有1 596个注释文本；自建脑血管疾病数据库来源于某省两医院的临床电子病历，总共1 062条文本信息。由于数据集规模小于百万级别，研究按照60:20:20的比例划分训练集、测试集和验证集，即两数据库60%的数据用于模型训练，20%的数据用于模型验证，20%的数据用于模型测试。实验数据集的划分如表1所示。

表1 实验数据集的划分

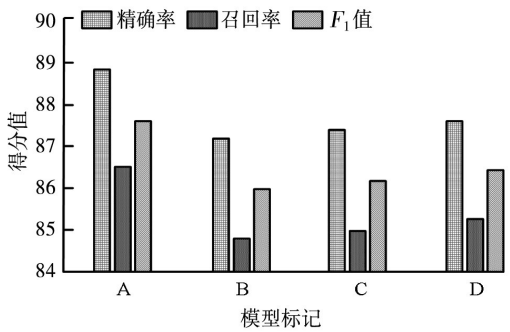
数据集	划分	症状体征	检查检验	疾病诊断	治疗	部位	实体总数
全国知识图谱与语义计算大会	训练集	6 085	9 925	6 996	10 939	17 183	17 920
	测试集	2 028	3 308	2 332	3 646	5 728	5 973
	验证集	2 028	1 985	2 332	3 646	5 728	5 973
自建脑血管疾病数据库	训练集	2 980	—	767	1 256	3 453	8 457
	测试集	993	—	256	419	1 151	2 819
	验证集	993	—	256	419	1 151	2 819

表1中，全国知识图谱与语义计算大会提供的医疗病历标注数据集中包含5类实体，分别是症状体征、检查检验、疾病诊断、治疗和身体部位，自建脑血管疾病数据库包含症状体征、疾病诊断、治疗和身体部位4类实体类型。融合多个特征的命名实体识别模型的参数设置如表2所示。

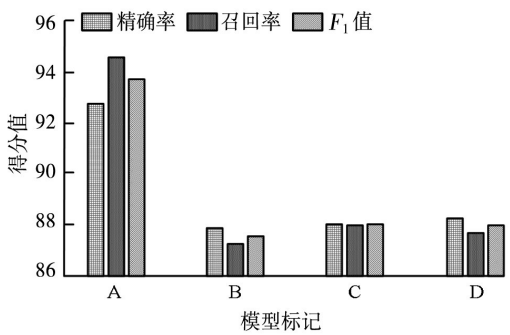
表2 融合多个特征的命名实体识别模型的参数设置

参数名称	设置值	参数名称	设置值
Learning rate	0.01	Batch_size	16
最大迭代次数	200	Dropout	0.4
字符嵌入维度	50	部首嵌入卷积神经网络窗口	3
词嵌入维度	50	优化器	Adam
部首嵌入维度	50	多头注意力机制	8
注意力层数量	6	—	—

为验证融合多个特征的命名实体识别模型(标记为A)的效果，实验将3种传统的命名实体识别模型作为对比，分别是Word2Vec-BiLSTM-CRF模型(标记为B)、Word2Vec-BiGRU-CRF模型(标记为C)以及IDCNN-CRF模型(标记为D)，并以精确率、召回率和F1值进行模型效果的评价，不同识别实体识别模型的评价结果如图3所示。



(a) 脑血管疾病数据集

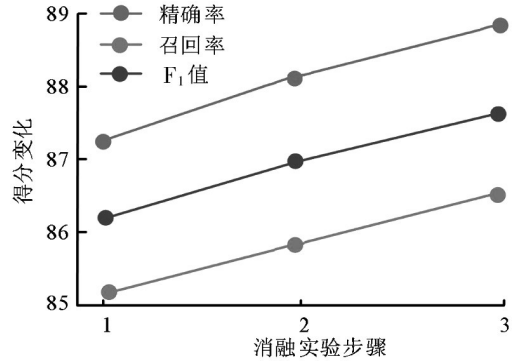


(b) 全国知识图谱与语义计算大会数据集

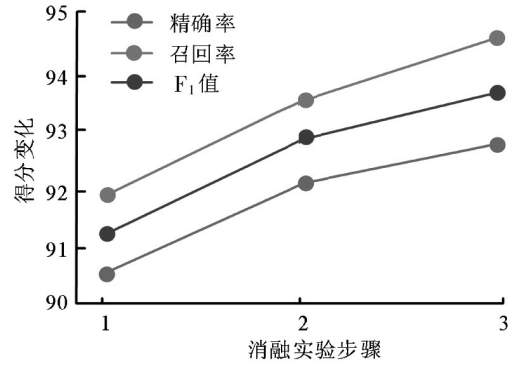
图3 不同识别实体识别模型的评价结果

图3(a)为脑血管疾病数据集中的模型评价结果，可以看到研究提出的融合多个特征的命名实体识别模型的精确率、召回率和F1值分别为88.86%、86.50%和87.63%，均高于三种传统的命名实体识别模型。图3(b)为全国知识图谱与语义计算大会数据集中的模型评价结果，可以

看到研究提出的融合多个特征的命名实体识别模型的精确率、召回率和F_1值分别为92.79%、94.65%和93.71%，均高于3种传统的命名实体识别模型。为验证融合多个特征的命名实体识别模型中各模块的效果，研究设计了消融实验的3个步骤，以交叉融合和拼接融合方式进行对比，消融实验结果如图4所示。



(a) 脑血管疾病数据集



(b) 全国知识图谱与语义计算大会数据集

图4 消融实验结果

图4中，步骤1为FLAT基线模型，步骤2为引入直接拼接，步骤3为引入交叉融合。可以看到，汉字部首特征的引入能够提升FLAT模型命名实体识别的效果，脑血管疾病数据集中模型精确率、召回率和F_1值分别提升了0.91%、0.73%和0.82%。全国知识图谱与语义计算大会数据集中模型精确率、召回率和F_1值分别提升了1.62%、1.65%和1.64%。融合多特征的命名实体识别模型的精确率、召回率和F_1值得到进一步提升，说明深层次的融合能够提升命名实体识别的效果。

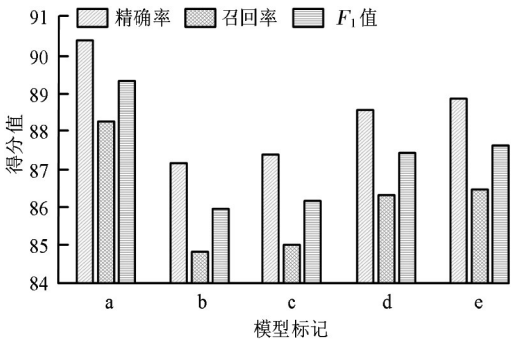
2.2 基于多任务学习的医疗病历命名实体模型实验分析

自建脑血管疾病数据集和全国知识图谱与语义计算大会数据集的实体实例和类型不同，两者又存在关联性，能够用于模型的并行训练。基于多任务学习的医疗病历命名实体模型的参数设置如表3所示。

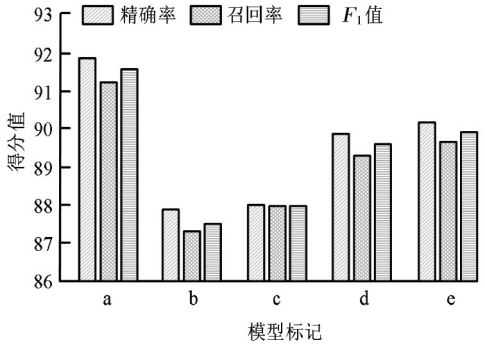
为验证研究提出的基于多任务学习的医疗病历命名实体模型(标记为a)的效果，实验将Word2Vec-BiLSTM-CRF模型(标记为b)、Word2Vec-BiGRU-CRF模型(标记为c)、BERT-BiLSTM-ATT-CRF模型(标记为d)和BERT-BiGRU-ATT-CRF模型(标记为e)作为对比，不同命名实体识别模型的实验结果如图5所示。

表3 基于多任务学习的医疗病历命名实体模型的参数设置

参数名称	设置值	参数名称	设置值
Learning rate	0.01	Batch_size	32
Dropout	0.06	Optimizer	Adam
Max_epoch	110	McGRU单元数	128
BiMcGRU层数	1	实体类型单元个数	6



(a) 脑血管疾病数据集



(b) 全国知识图谱与语义计算大会数据集

图5 不同命名实体识别模型的实验结果

图5(a)为脑血管疾病数据集中的模型实验结果，可以看到研究提出的基于多任务学习的医疗病历命名实体模型的精确率、召回率和F_1值分别为90.39%、88.30%和89.34%，均高于其他4种传统的命名实体识别模型。图5(b)为全国知识图谱与语义计算大会数据集中的模型实验结果，可以看到研究提出的基于多任务学习的医疗病历命名实体模型的精确率、召回率和F_1值分别为91.86%、91.20%和91.53%，均高于其他四种传统的命名实体识别模型。为进一步验证BiMcGRU模型的效果，实验将BiGRU模型作为对比，不同模型的识别结果如表4所示。

表1中，下标1表示脑血管疾病数据集中的识别结果，下标2表示全国知识图谱与语义计算大会数据集中的识别结果。单任务训练下，BERT-BiMcGRU-ATT-CRF模型

(下转第130页)