基于字形特征的中文医学命名实体识别方法

孟伟伦[1,2],郭景峰[1,2*],邢珂萱[1],魏 宁[1,2],王巧梭[3],刘 滨[4]
(1. 燕山大学信息科学与工程学院,河北秦皇岛 066004;2. 河北省虚拟技术与系统集成重点实验室,河北秦皇岛 066004;
3. 河北建材职业技术学院,河北秦皇岛 066000;4. 河北科技大学大数据与社会计算研究中心,河北石家庄 050018)

摘 要: 作为医学信息抽取的第一个关键环节,医学命名实体识别任务旨在从如电子医疗病例、中文医药说明书等非结构化文本中抽取出医学相关的实体. 目前大多数中文医学命名实体识别工作通过在预训练模型上进行微调来获得文本表示向量,然后利用特征工程来提升模型在医疗领域上的性能. 这些模型大部分源自在通用数据集上表现较好的模型,没有考虑中文医学数据集的语言特性. 通过在多个医学数据集上进行统计分析,发现部分类型的医学实体在字形上具有共性,如在汉字中大部分表示疾病含义的字符都包含“疒”,大部分表示身体器官的字符都包含“月”. 针对这些问题,本文提出了一种基于字形特征的中文医学命名实体识别方法,该方法通过在文本表示向量上融合字形向量以及进一步利用数据集中负样本来提升模型的准确度和泛化能力. 在多个公共的中文医学数据集上的实验结果表明,该方法获得了比其他模型更好的效果,并且通过消融实验证明了融合字形特征和从负样本中学习对于该任务是有效的.

关键词: 字形;负样本;两阶段;医学信息;命名实体识别;深度学习
基金项目: 河北省省级科技计划(No.21310101D);中央引导地方科技发展资金(No.226Z0102G);国家文化和旅游科技创新工程(2020年度)
中图分类号: TP391.1;TP183 **文献标识码:** A **文章编号:** 0372-2112(2024)06-1945-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230516

A Chinese Medical Named Entity Recognition Method Based on Glyph Features

MENG Wei-lun[1,2], GUO Jing-feng[1,2*], XING Ke-xuan[1], WEI Ning[1,2], WANG Qiao-suo[3], LIU Bin[4]
(1. *School of Information Science and Technology, Yanshan University, Qinhuangdao, Hebei* 066004, *China*;
2. *The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao, Hebei* 066004, *China*;
3. *Hebei Construction Material Vocasional and Technical College, Qinhuangdao, Hebei* 066000, *China*;
4. *Big data and Social Computing Research Center, Hebei University of Science and Technology, Shijiazhuang, Hebei* 050018, *China*)

Abstract: As the first key link in medical information extraction, the medical named entity recognition task aims to extract medical-related entities from unstructured texts such as electronic medical records and Chinese medical instructions. Most current Chinese medical named entity recognition works obtain text representation vectors by fine-tuning pre-trained models, and then use feature engineering to improve the performance of the models in the medical field. Most of these models are derived from models that perform well on general-purpose datasets, without considering the language characteristics of Chinese medical datasets. Through statistical analysis on multiple medical data sets, it is found that some types of medical entities have similarities in glyphs. For example, in Chinese characters, most of the characters representing diseases contain “疒”, and most of the characters representing body organs contain “月”. In response to these problems, this paper proposes a Chinese medical named entity recognition method based on glyph features. This method improves the accuracy and generalization ability of the model by fusing the glyph vector on the text representation vector and further utilizing the negative samples in the dataset. Experimental results on multiple public Chinese medical datasets show that this method achieves better results than other models, and ablation experiments prove that fusing glyph features and learning from negative samples is effective for this task.

收稿日期:2023-06-08;修回日期:2024-01-17;责任编辑:李勇锋
*通讯作者:郭景峰

Key words: glyph feature; negative sample; two stages; medical information; named entity recognition; deep learning
Foundation Item(s): S&T Program of Hebei (No.21310101D, No.226Z0102G); National Cultural and Tourism Science and Technology Innovation Project (2020)

1 引言

作为自然语言处理的一项基本任务,命名实体识别(Named Entity Recognition,NER)旨在抽取文本中具有特定意义的实体同时将实体的属类标注出来[1],在医疗信息中常见的实体如疾病、手术、药物等,这些实体能够为后续的下游任务如关系抽取、医疗信息问答等提供更准确的信息[2].

目前医学领域的命名实体识别方法大多采用在预训练模型上进行微调来获得文本表示向量,这些预训练模型如Bert、Robert等都是基于通用语料训练的,并且在微调过程中受限于训练集的规模较小,导致获得的文本向量无法充分地包含医学领域上的动态语义信息,因此需要引入更多的特征信息来丰富文本向量表示[3]. 目前在中文NER任务上提出的方法主要是通过引入外部知识来实现,如词典、知识图谱、搜索引擎等[4,5],它们都取得了不错的效果,但需要人工构建,耗时耗力,且最终性能受外部知识的影响较大,泛化能力较弱,并且目前缺乏在医学领域较突出的外部知识,这些都影响了医学NER任务的发展.

在英语中不同的单词可能具有相同的词根或词缀以更好地表示语义,例如:“biology”、“biotech”等表示生物学相关的词包含相同的前缀“bio-”,这有助于识别某类型或主题的实体. 对中文而言,也有类似的结构信息,在医学语料中,例如:肝、肺、肾脏等与身体器官有关的词一般包含“月”字形,痛风、瘟疫等疾病相关的词一般包含“疒”字形. 本文对比了两种汉字拆解方式,如表1所示.

表1 汉字拆解方式

汉字	MD	SD
肺	月、市	冂、巾、丨、乙、一
癌	疒、品、山	冂、一、乙、丶、氵、山、口、凵、丨、丿、丬

汉字由象形文字演变而来,其结构在一定程度上可以反映所包含的语义信息. 在多个公开中文医学数据集上进行拆字后,统计分析发现不同类别的医学实体在字形上存在明显的差异性,并且在字形分布上与其对应实体类型的解释(标注规范)之间具有相关性. 因此,对于中文医学NER而言,在模型中融入字形结构特征来丰富文本向量表示应是可行的.

为解决该问题,对上述数据集的字形统计结果继续分析,发现这些数据集在负样本上存在问题. 不同数据集的负样本在字形上存在很大相似性,同时在其占比最多的前15种字形上都出现了“月”、“疒”这些与医学实体相关的字形. 同时,参考Yangming Li等人在有关负采样上的研究发现在许多情况下,NER模型严重受到未标记实体问题的困扰[6,7]. 其原因是:模型在学习实体类型的特征时被负样本中存在的部分未标记实体所误导,影响了模型效果. 对此,构建了一个简单的负样本学习器来找出这些未标记实体,减轻它们对模型的影响.

同时,针对医学NER任务中常见的实体嵌套问题,本文选择基于指针网络的NER方法作为基础模型,并在此基础上提出了一种基于字形特征的中文医学命名实体识别方法(G-CMNER),并在多个公开中文医学数据集上进行实验,以评估该方法的有效性. 结果表明,G-CMNER在所有数据集上都取得了最好的结果.

2 相关工作

近年来,随着深度学习技术的不断发展,基于深度学习的命名实体识别方法获得了较多的关注. 在医学领域中,Ling等人提出一种基于领域知识增强的LSTM-CRF模型[8],将领域知识编码成词典嵌入和外部标签嵌入,作为模型输入的一部分,在NCBI疾病语料库上证明了领域知识对医学NER任务的有效性. Li等人提出基于动态注意力机制的方法,构造一个联合嵌入单元来获取领域知识,领域知识由领域数据集训练的词向量提供[9]. 这些方法主要是通过引入外部知识来提高词向量质量. 此外,研究者们也提出了一些利用中文内部结构信息的方法. Dong等人首次提出在中文NER模型中使用字形信息,使用Bi-LSTM提取字形嵌入,将其与字符嵌入拼接起来使用[10]. 崔少国等人于2022年提出一种融合语义及边界信息的中文电子病历NER模型(WHSemantic+Lattice),该方法将汉字图像特征和五笔字型编码进行融合作为高级语义信息用于中文电子病历数据[11]. 上述这些方法均是通过有限的外部知识来提升词向量的质量,没有使用预训练模型.

最近,基于大规模无标注数据的预训练模型(如BERT[12]、GPT[13]等)取得了巨大成功,很多基于BERT的扩展研究在医学NER任务上都表现得非常出色. Yingjie Gu等人2022年提出RICON模型,用于研究中文NER中实体跨度的规律性,包括规律感知模块和规律诊断模块,其在基准数据集和医疗数据集上优于以前的最先进方法[14]. 吴炳潮等人2022年提出一种动态迁移实体块信息的跨领域中文NER模型(TES-NER),该方法将

跨领域共享的实体块信息通过基于门机制的动态融合层,从语料充足的通用领域动态迁移到垂直领域,在医学领域的数据集上取得了较好的效果[15]. 虽然这些基于预训练模型的方法已经取得了不错的效果,但是由于预训练模型大多是基于通用领域的文本语料训练的,无法更多地包含医学领域的上下文语义信息. 对此,受NER任务中利用中文内部结构信息研究工作的启发,本文提出了一种基于字形特征嵌入的中文医学命名实体识别方法,将由预训练模型BERT得到的字符特征与字形特征相结合,以达到更有效地识别医学实体的目的.

3　数据分析

为了解不同类型的医学实体之间在字形结构分布上是否具有明显的差异性,以及该特征在相应类型的实体上是否普遍存在,本章节在5.1节介绍的多个公开数据集上进行了数据统计和分析.

3.1　准备工作

3.1.1　汉字结构拆解方法

汉字按结构可以划分为独体字和合体字两种,独体字指汉字结构中仅含有一个单独形体、不可拆分为两个或两个以上形体的字,与合体字的概念相对. 如表1所示,本文将汉字结构的拆解分为独体字拆解(Monolithic Dismantling,MD)和笔画拆解(Stroke Dismantling,SD)两种方式. MD方式会将合体字中的独体字部分完整保留;而SD方式则不保留合体字中的独体字部分,全部按笔画拆解. 这两种方式均按照汉语拆字词典提供的拆字字典数据库进行汉字拆解. 在5.1节中对这两种方式在模型上的效果进行了对比实验,结果表明基于MD方式的模型效果更好,因此,这里仅展示MD方式的拆解结果.

3.1.2　统计指标

统计时发现字形结构存在长尾分布现象,在三个数据集的总计20类实体中,前15种字形的总和占据了总数的40%~50%,但这些字形仅占全部拆分字形种类数的3%~5%. 因此,这里只记录每种类型的实体中数量最多的前15种字形.

本文还设计了两种指标用以衡量字形结构特征的影响:类型占有率K_p和实体占有率E_p.

$$K_p=\frac{R_s}{R_e},E_p=\frac{R_s}{E_s} \tag{1}$$

其中,R_s、R_e分别为当前实体类型的某一字形的数量和字形总数,E_s为当前类型的实体总数. K_p用于衡量不同类型的实体中字形的分布情况,E_p用于衡量该字形在当前实体类型中是否普遍存在,E_p越高,表示该字形与当前实体类型的关联程度越高.

3.2　总体分析

从图1可以看出,不同类别的医学实体之间在字形结构的种类和分布上具有明显的差异性. 结合表2来看,这种差异性与实体类型描述之间也存在关联. 但是在负样本的字形构成上也发现了这些与医学实体相关的字形且占比相对较高,该问题在三个数据集上均存在,在一定程

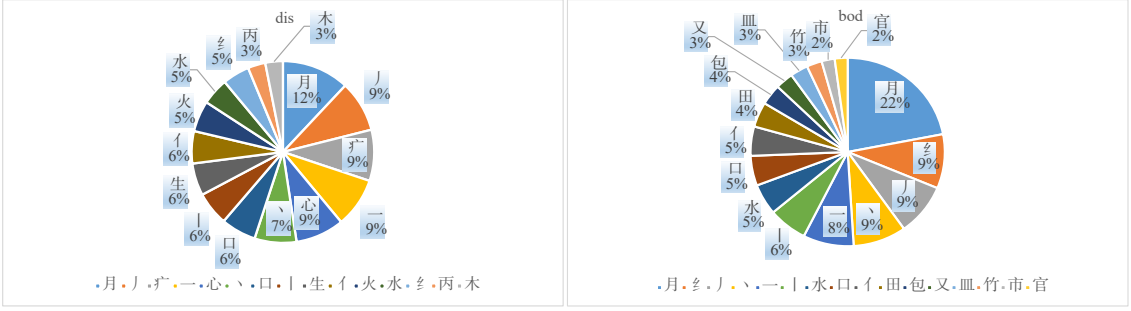


图1　CMeEE中部分实体类型的字形类型占比图

表2　CMeEE数据集中部分实体类型和非实体的字形统计以及实体类型描述

实体类型	占比最多的前15种字形	解释
疾病(dis)	月、丿、疒、一、心、丶、口、丨、生、亻、火、水、纟、丙、木	指导致病人处于非健康状态的原因或者医生对病人做出的诊断,包括:疾病或综合征、中毒或受伤、器官或细胞受损
身体(bod)	月、纟、丿、丶、一、丨、水、口、亻、田、包、又、皿、竹、市	泛指细胞、组织、及位于人体特定区域的由细小物质成分组合而成的结构、器官、系统、肢体,另外包括身体产生或解剖身体产生的物质等,主要包括:身体部位、身体物质
微生物(mic)	艹、困、疒、丙、丰、毋、水、一、月、亻、口、丿、纟、干、丶	微生物类包括细菌、病毒、真菌以及一些小型的原生生物、 显微藻类等在内的一大类生物群体,另外包括微生物类产生的毒素、激素、酶等
负样本	一、丿、口、亻、丶、丨、月、日、白、勺、土、丁、木、水、心	—

度上可以反映出医学NER数据集的负样本中存在部分未标记实体.

从表3中可以看出,疾病类型(dis)的主要字形特征如“月”、“疒”和“心”的字形占有率都超过了35%,该数据表明每三个疾病实体中至少会有一个实体包含这些表征类型的字形. 类型占有率和实体占有率都较高的字形表明其与对应实体类型之间的关联程度越高,对模型的提升能力越大.

结合表3和图2可以看出不同数据集的同一类型实体在字形分布上是相似的,一定程度上可以表明这种关联程度在医学数据集中是普遍存在的,字形特征的引入应该可以提升模型的泛化能力和稳健性.

结合对字形类型占有率和实体占有率统计结果的分析,设计了一个用于反映数据集字形表征程度的度量指标:

$$\mathrm{DDGR}=\sum_{i=1}^{n}\left(K_E f(E_P)\right) \tag{2}$$

$$f\left(E_P\right)=\begin{cases}\max\left(E_P(1),E_P(2),\cdots,E_P(k)\right),\dfrac{\left(E_P(k)-E_P(1)\right)}{k}\leqslant 0.05\\ \max\left(E_P(2),E_P(3),\cdots,E_P(k-1)\right),0.05<\dfrac{\left(E_P(k)-E_P(1)\right)}{k}\leqslant 0.08\\ \dfrac{\sum_{i=2}^{n}E_P(i)}{n-2},\dfrac{\left(E_P(k)-E_P(1)\right)}{k}>0.08\end{cases} \tag{3}$$

其中,K_E代表每类实体占实体总数的比例,$f(E_p)$代表该实体类型的平均实体占有率. 利用该公式对上述三个数据集的字形表征分数进行了计算,结果为:CMeEE>CCKS2019>CCKS2017,分数越高,代表该数据集的字形特征越明显,模型的效果应该越好,后面的5.5节对此进行了实验分析.

表3 CMeEE数据集的部分字形实体占有率统计图

dis		bod		mic	
char	E_p	char	E_p	char	E_p
月	52.5	月	63.8	⺾	48.1
丿	37.4	纟	27.1	困	42.2
疒	37.3	丿	25.9	疒	34.9
一	36.4	丶	25.7	丙	29.3
心	35.5	一	24.6	丰	29.1
丶	30.2	丨	18.8	毋	26.9
口	25.6	水	15.4	水	20.7
丨	24.3	口	14.9	一	20.5
生	23.6	亻	14.6	月	15.3
亻	23.2	田	12.4	亻	14.7

4 模型框架

4.1 相关定义

为了更好地描述框架和提出的模型,这里给出了如下定义:

定义1 跨度:在命名实体识别中,跨度为命名实体在文本中的字符位置范围,通常表示为一个(起始位置,结束位置)二元组. 对于句子X中的某个跨度可以表示为$s=\left(x_i,x_j\right)$.

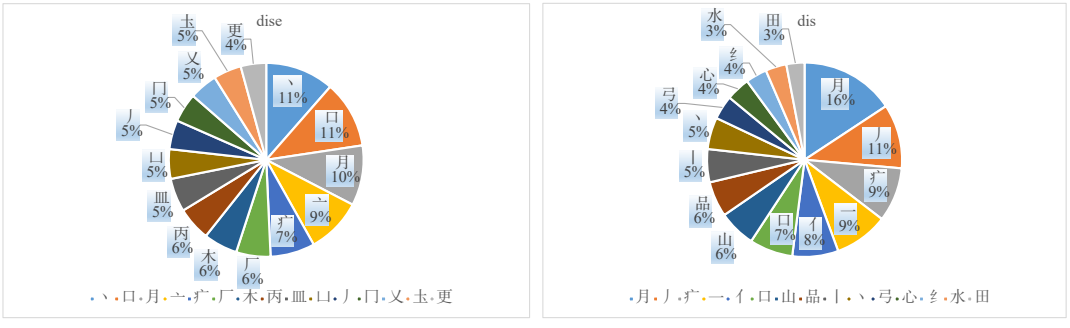
定义2 基础字形:对于一个汉字$z\in Z$,其中Z表示所有汉字的集合,按照MD方式拆解得到的汉字组件称为基础字形.

定义3 字形集:所有汉字的基础字形组成的集合称为字形集.

定义4 拆字字典:所有汉字及其对应的字形集,叫做拆字字典.

定义5 字形向量:将字形集中的所有基础字形按照一定顺序排列得到一组数字,这组数字称为字形向量.

定义6 字形向量查找表:字形集中的所有基础字



CCKS2017 CCKS2019

图2 三个数据集中疾病类型实体的字形分布图

形及其对应的字形向量共同组成了字形向量查找表.

4.2 基于跨度预测的NER框架

基于跨度预测的NER框架通常由三个主要模块构成:文本表示层、跨度表示层和跨度预测层.

4.2.1 文本表示层

给定一个句子 $X=(x_1,x_2,\cdots,x_n)$,其中 n 为句子的长度,将 X 输入到文本表示层,通过上下文化词嵌入得到句子的字符向量表示:

$$\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)=E\text{mbedding}(x_1,x_2,\cdots,x_n) \tag{4}$$

其中,Embedding() 使用预训练模型 BERT 作为模型的上下文化词嵌入.

4.2.2 跨度表示层

对于句子 X,通过枚举所有可能的种子跨度 $S=(s_0,s_1,\cdots,s_k)$,然后为每个跨度 s_i 重新分配一个标签 $\boldsymbol{y}\in\boldsymbol{Y}$,其中,$\boldsymbol{Y}$ 为标签集. 对于跨度 $s_i=(x_{b_i},x_{e_i})$,$1\leqslant b_i\leqslant e_i\leqslant n$,其中,$b_i$ 和 e_i 分别表示跨度的左、右边界,其最终向量表示是通过连接实体的左右边界来计算的:$\boldsymbol{t}_i=[\boldsymbol{h}_{b_i};\boldsymbol{h}_{e_i}]$.

4.2.3 跨度预测层

跨度的向量表示 $\boldsymbol{t}_i$ 被输入到 softmax 分类器中计算该跨度属于标签 $\boldsymbol{y}$ 的概率.

$$P(\boldsymbol{y}|\boldsymbol{t}_i)=\frac{\text{score}(\boldsymbol{t}_i,\boldsymbol{y})}{\sum_{\boldsymbol{y}'\in\boldsymbol{Y}}\text{score}(\boldsymbol{t}_i,\boldsymbol{y}')} \tag{5}$$

其中,score(·)是用于衡量跨度和标签之间兼容性的函数:

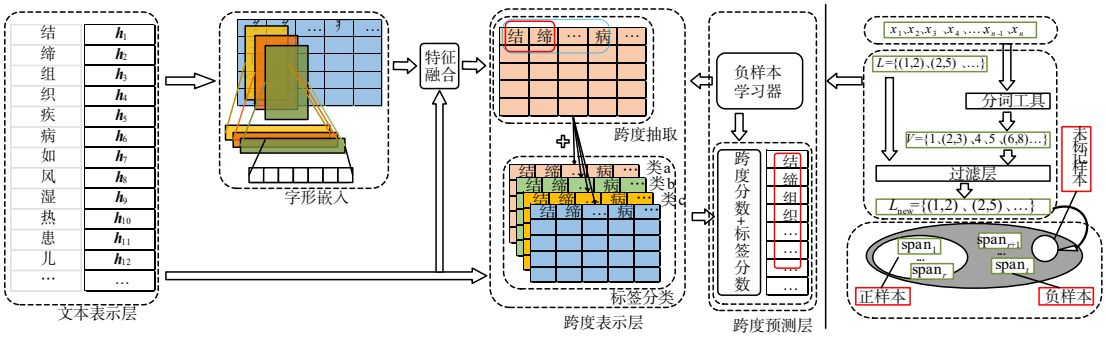
$$\text{score}(\boldsymbol{t}_i,\boldsymbol{y}_k)=\exp(\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{y}_k) \tag{6}$$

其中,$\boldsymbol{y}_k$ 是标签 k 的可学习表示.

5 嵌入字形的NER方法

为解决使用基于通用数据集的预训练模型无法更多地包含医学领域的动态语义信息问题,在基于跨度预测的NER框架下,使用膨胀卷积神经网络ID-CNN作为字形嵌入层,将预训练模型BERT得到的字符向量与字形向量相融合,以增强模型的文本向量表示. 同时对于未标记实体问题,考虑到这些实体可能会在字形特征上误导模型,构建了一个负样本学习器用于从负样本中抽取这些未标记实体.

模型的整体结构如图3所示,首先将由字形嵌入模块获得的字形向量与字符向量相融合,得到的融合向量输入到跨度抽取模块获得可能的种子跨度,然后在标签分类模块中使用原有的字符向量计算候选的种子跨度的不同类别上的标签得分,最后将跨度抽取模块的跨度得分与标签得分相加得到实体得分,根据该得分找出待抽取的实体及其所属的类别.



(a) 模型整体结构图　　(b) 负样本学习器

图3　G-CMNER模型

5.1 字形嵌入

给定一个输入句子 X,首先通过拆字字典 P 得到句子中每个字的字形序列 $x_i=(k_{i_1},k_{i2},\cdots,k_{im})$,然后使用字形向量查找表 R 获得 x_i 的字形向量表示 $\boldsymbol{w}_i=(\boldsymbol{w}_{i_1},\boldsymbol{w}_{i_2},\cdots,\boldsymbol{w}_{i_m})$:

$$(k_{i_1},k_{i2},\cdots,k_{\text{im}})=P(x_i) \tag{7}$$

$$(\boldsymbol{w}_{i_1},\boldsymbol{w}_{i_2},\cdots,\boldsymbol{w}_{i_m})=R(k_{i_1},k_{i2},\cdots,k_{\text{im}}) \tag{8}$$

最后句子的字形嵌入表示为 $\boldsymbol{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n)$.

使用膨胀卷积神经网络对句子文本经过字形嵌入得到的字形向量序列进行特征提取,相比于普通的卷积神经网络优势在于在不做pooling层损失信息的情况下,加大了感受野,使得每个卷积输入都能包含较大范围的信息,更适合于长文本序列.

5.2 两阶段抽取和多头识别

在G-CMNER方法中使用改进的两阶段抽取方式取代了之前的跨度预测层,改进包括三个方面:一是在跨度抽取和标签分类模块上分别引入实体得分函数、标签得分函数,并将它们的和作为最终的实体标签得分,以减少错误传播的影响;二是标签分类模块

使用全量样本，而不是经跨度抽取模块筛选后的样本，以提高标签分类模块的效果；三是为降低未标记实体对相应实体类型的影响，采取与Peng等人在正向无标记学习模型中一样的策略，构建多头标签识别器[16].

与跨度抽取模块不同，考虑到字符向量和字形向量来自不同的向量空间，融合后的向量可能会对标签分类造成影响，这里使用的是经预训练模型获得的字符向量作为输入，因为该字符向量在语义空间上更容易解释字符之间的相关性.

5.2.1 跨度抽取模块

将句子X的字符向量$\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)$和字形向量$\boldsymbol{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n)$连接起来，输入到一个全连接层进行信息融合：

$$\boldsymbol{c}_i=(\boldsymbol{h}_i\oplus\boldsymbol{w}_i)\boldsymbol{W}^{\circ}+\boldsymbol{b} \tag{9}$$

其中，$\oplus$表示连接操作，$\boldsymbol{W}^{\circ}$和$\boldsymbol{b}$都是可学习的参数.

然后种子跨度$\boldsymbol{s}_i$的跨度向量可以表示为$(\boldsymbol{c}_i,\boldsymbol{c}_j)$，分别经过头、尾变换矩阵得到跨度的头尾表示向量$\boldsymbol{q}_i$、$\boldsymbol{k}_j$：

$$\boldsymbol{q}_i=\boldsymbol{w}_{q,i}\boldsymbol{c}_i \tag{10}$$

$$\boldsymbol{k}_j=\boldsymbol{w}_{k,j}\boldsymbol{c}_j \tag{11}$$

其中，$\boldsymbol{w}_{q,i}\in\boldsymbol{W}_q$和$\boldsymbol{w}_{k,j}\in\boldsymbol{W}_k$是头尾变换矩阵.

使用跨度$\boldsymbol{s}_i$经变换后的头尾向量得到的该跨度的跨度得分，跨度得分函数为

$$\mathrm{Score}_S(i,j)=\boldsymbol{q}_i^{\mathrm{T}}\odot\boldsymbol{k}_j \tag{12}$$

其中，$\odot$表示向量的内积操作.

5.2.2 标签分类模块

标签分类模块用于预测跨度$\boldsymbol{s}_i$属于标签集$\boldsymbol{Y}$的所有可能得分. 句子$\boldsymbol{X}$的字符向量为$\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)$，此时$\boldsymbol{s}_i$可以表示为$(\boldsymbol{h}_i,\boldsymbol{h}_j)$，然后计算其标签得分：

$$\mathrm{Score}_L(i,j)=\boldsymbol{W}_L^{\mathrm{T}}\left[\boldsymbol{h}_i:\boldsymbol{h}_j\right] \tag{13}$$

5.2.3 实体抽取

该模型使用跨度得分和标签得分相加作为最终的实体跨度得分：

$$\mathrm{Score}=\mathrm{Score}_S+\mathrm{Score}_L \tag{14}$$

对于跨度$\boldsymbol{s}_i$，其最终的打分函数可以表示为

$$\mathrm{Score}=\boldsymbol{w}_{q,i}\boldsymbol{w}_{k,j}(\boldsymbol{h}_i\oplus\boldsymbol{w}_i)(\boldsymbol{h}_j\oplus\boldsymbol{w}_j)+\boldsymbol{W}_L^{\mathrm{T}}\left[\boldsymbol{h}_i:\boldsymbol{h}_j\right] \tag{15}$$

然后计算该跨度属于标签y的概率：

$$P(\boldsymbol{y}|\boldsymbol{s}_i)=\frac{\mathrm{Score}(\boldsymbol{s}_i,\boldsymbol{y})}{\sum_{\boldsymbol{y}'\in\boldsymbol{Y}}\mathrm{Score}(\boldsymbol{s}_i,\boldsymbol{y}')} \tag{16}$$

5.3 负样本学习器

通过对医学数据集中的实体进行分析，发现不同于通用领域的数据集，医学实体的跨度相对较长，同时这些长实体可以通过拆分得到的部分短实体也可能是医学实体. 对于医学数据集中的这一特点，可以设计相应的规则从整个数据集中挖掘到未标记实体.

如图3(*b*)所示，给定一个句子X和它的标注序列$L=(l_1,l_2,\cdots,l_n)$，由该序列得到实体集合Span，同时引入经医学数据集调整后的分词工具对句子进行分词，得到词组集合V. 按照实体跨度的大小，可以将Span看作粗粒度，V看作细粒度，合并Span和V，选取其中较小粒度的词组，得到更新后的词组集合. 然后对其中的词汇按照一定的规则进行组合，获得新的标注序列V_{new}. 负样本学习器将该标注序列加入到模型中一起训练，既能减少未标记实体对模型的误导，又使得模型可以发现更多的潜在实体，以提升模型的泛化能力和稳健性.

原本的损失函数L_{base}为

$$\log\left(1+\sum_{(i,j)\in P}\mathrm{e}^{-s(i,j)}\right)+\log\left(1+\sum_{(i,j)\in Q}\mathrm{e}^{s(i,j)}\right) \tag{17}$$

增加的未标记实体部分的损失函数L_{neg}为

$$\log\left(1+\sum_{(i,j)\in \bar{P}}\mathrm{e}^{-s(i,j)}\right)+\log\left(1+\sum_{(i,j)\in \bar{Q}}\mathrm{e}^{s(i,j)}\right) \tag{18}$$

更改后总的损失函数为

$$L_{\mathrm{total}}=\lambda_1 L_{\mathrm{base}}+\lambda_2 L_{\mathrm{neg}} \tag{19}$$

其中，原始标注序列的损失和未标记实体部分的损失分别由λ_1和λ_2加权.

6 实验分析

6.1 实验数据集

为检验所提出方法的有效性，本文在CCKS2017、CCKS2019和CMeEE三个公开数据集上进行了评估，其中CCKS2017和CCKS2019分别为2017年和2019年全国知识图谱与语义大会的评测任务数据，CMeEE数据集来自阿里天池平台承办的中文医疗信息处理挑战榜CBLUE，是目前国内比较优秀的公开评测基准，关于这三个数据集的详细信息见表4和表5.

表4 三个实验数据集的语料统计信息

数据集		实体总数	不同长度的实体占比/%		句子数量
			0<实体长度<5	5<实体长度<10	
CCKS2017	训练集	53 801	82.72	15.93	2 007
	测试集	5 265	85.49	12.92	223
CCKS2019	训练集	17 653	65.97	24.39	1 000
	测试集	6 002	75.51	19.83	379
CMeEE	训练集	61 796	61.78	30.43	15 000
	训练集	53 801	82.72	15.93	2 007

表5　三个实验数据集的命名实体类型详情

数据集	类别数	命名实体类型
CCKS2017	5类	身体部位(body)、症状和体征(symp)、疾病和诊断(dise)、检查和检验(chec)、治疗(cure)
CCKS2019	6类	疾病和诊断(dis)、手术(pro)、解剖部位(anatomy)、药物(drug)、影像检查(testimage)、实验室检验(testlab)
CMeEE	9类	疾病(dis)、身体(bod)、临床表现(sym)、药物(dru)、微生物(mic)、医疗设备(equ)、医疗程序(pro)、医学检验项目(ite)、科室(dep)

6.2　评价指标

目前NER任务上采用的评价指标主要包括3个:精确率P(Precision)、召回率R(Recall)和$F1$值.在这里,本文使用的是实体级别的评价指标,以实体为计算单位,单个实体中全部字符的标签都正确才算正确.同时,由于损失函数和评价指标都是以实体为单位的,这也保证了训练、评估和预测各阶段的一致性.

6.3　基线方法

目前,医学领域的中文命名实体识别任务的相关数据集和模型大多来源于相关的比赛或会议,这些模型大多没有开源,且使用的数据集也不尽相同,因此无法直接使用这些模型的评测结果进行比较.对此,本文选取了包括FLAT(base)[17]、FLAT(bert)、MECT[18]、BERT-CRF、BERT-Softmax、BERT-Span[19,20]、Global-Pointer、RICON、WHSemantic+Lattice和TES-NER共10个模型作为基线,其中前7个模型为近三年来在通用领域上表现较好的模型,这些模型复现后均在医学数据集上进行了调整,以获得更好的结果.最后的三个模型为近两年在中文医学领域数据集上较出色的模型.

6.4　实验分析

本文提出的G-CMNER方法与基线方法在CCKS2017、CCKS2019、CMeEE三个公开中文医学NER数据集上的实验结果如下.

如表6所示,在上述三个数据集上进行的对比实验可以分为3部分:第1部分的基线方法没有使用预训练模型,第2部分的基线方法均使用了预训练模型,第3部分为本文提出的G-CMNER模型.从实验结果中可以看出:(1)在CCKS2017和CCKS2019两个数据集上,第2部分的部分模型在性能上要弱于第1部分中的部分模型,表明在中文医学NER任务上使用预训练模型的方法在性能上并不总好于未使用的方法,这在一定程度上反映了基于通用数据集得到的预训练模型在该任务上存在语义表征能力不足,缺乏医学领域中特有的语义信息的问题.而第3部分中的G-CMNER模型针对该问题,利用汉字在医学实体中的字形特点,使用汉字的字形结构信息作为补充,在上述三个数据集上均获得了最好的效果,其在CCKS2019和CMeEE数据集上相比于基线中的最好结果在$F1$指标上分别提升了1.18%和2.34%,这表明汉字的字形结构信息以及在医学实体上表现的规律性对于中文医学NER任务来说是有效的,进一步提升了模型的性能.同时该方法在三个数据集上的提升效果与前面3.2节中得到的数据集字形表征分数的排序是一致的,也证明了该指标的有效性.(2)从精确率P和召回率R两个方面再次对实验结果进行分析,可以看出G-CMNER模型在提高精确率P的同时,在召回率R上也有明显的提升,结合表2中关于三个数据集中不同长度实体的占比情况来看,实体长度在5~10范围的长实体占比越大,模型在召回率R指标

表6　G-CMNER方法与基线模型在三个公开中文医学NER数据集上进行比较　　单位:%

		CCKS2017			CCKS2019			CMeEE		
		P	R	$F1$	P	R	$F1$	P	R	$F1$
模型	FLAT(base)	**92.59**	91.66	92.13	82.08	79.91	80.98	59.84	60.05	59.94
	MECT	92.44	91.51	91.98	81.45	82.76	82.11	60.36	60.38	60.37
	WHSemantic+Lattice	91.7	90.95	91.33	**84.31**	81.83	83.07	62.79	61.39	62.09
	TES-NER	90.29	91.35	90.82	81.68	82.22	81.95	62.53	62.83	62.68
使用预训练模型	FLAT(bert)	90.42	92.37	91.39	81.44	**83.19**	82.30	56.14	47.63	51.33
	BERT-CRF	84.30	67.21	74.79	72.64	67.90	70.19	66.66	64.93	65.78
	BERT-Softmax	80.87	63.17	70.93	70.96	67.26	69.06	63.33	62.78	63.05
	BERT-Span	78.44	56.37	65.60	75.05	66.54	70.54	**66.90**	63.51	65.16
	GlobalPointer	86.43	89.63	87.67	83.35	82.26	82.81	65.35	**66.61**	**65.98**
	RICON	92.33	91.99	92.16	83.47	82.89	**83.18**	66.25	64.89	65.57
	TES-NER(bert)	91.3	**93.4**	92.34	81.97	83.1	82.54	64.06	65.43	64.75
	G-CMNER	**92.65**	**92.36**	92.51	**85.61**	**83.28**	**84.36**	**67.76**	**68.88**	**68.32**

上的提升效果就越明显，说明模型的负样本学习模块能够充分利用医学NER中长实体较多的特点帮助模型发现更多的潜在实体，进一步提升模型的性能．综上所述，相比于基线方法，G-CMNER模型在融合字形结构信息与引入负样本学习器后有效地提升了在中文医学NER任务上的识别性能．

如表7所示，针对前面3.1节中提到的两种汉字拆解方法，本文在CCKS2017、CCKS2019和CMeEE三个数据集上进行了对比实验，这里并没有对模型进行参数调优，只是在保持参数不变的情况下，对比两种方法的效果．从实验结果中可以看出，MD方式确实要好于SD方式，这可能是由于基于SD方式得到的字形中会包含很多如“一”、“丨”、“丿”这样的基本字形，这些字形会造成不相关的字符之间在字形上也存在共性，对模型造成误导，影响了模型的性能．

表7 汉字结构拆解方法的对比实验

数据集	模型	*P*/%	*R*/%	*F*1/%
CCKS2017	G_MD	85.94	88.9	87.42
	G_SD	85.26	88.6	86.93
CCKS2019	G_MD	85.96	82.56	84.26
	G_SD	84.27	84.11	84.19
CMeEE	G_MD	67.22	67.84	67.53
	G_SD	66.95	66.73	66.84

如表8所示，针对字形特征向量的融合尝试了多种方法，包括：向量拼接(cat)、向量相加(add)、注意力网络(使用相对位置编码)+向量拼接(att+cat)、注意力网络(使用相对位置编码)+向量相加(att+add)，本文在CCKS2017、CCKS2019和CMeEE三个数据集上对这些方法进行了对比实验．从实验结果中可以看出：上述四种融合方式中，向量拼接的效果最好，使用自注意力机制的模型(att+cat)和模型(att+add)的性能下降最明显．对于向量拼接(cat)和向量相加(add)两种方式之间的性能差异，可能是由于在使用向量拼接(cat)方法进行融合时加入了全连接层，这会使得属于不同向量空间中的字符向量与字形向量之间产生交互，并且该交互对于不同向量空间之间的向量融合是有效的．而对于使用自注意力机制进行向量融合带来的性能下降问题，可能是由于字形向量特征在构建过程中没有考虑到上下文关系，若直接使用注意力网络会学习到错误的上下文信息．

此外，由于预训练模型本身具有的较大参数量，使得本文提出的G-CMNER模型在时间复杂度上要高于未使用预训练模型的方法．同时相比于其它使用预训练模型但未引入外部信息的方法，由于引入了字形结构信息和负样本学习器两个模块，导致模型的参数量增加，使得运行时间也相应增加．对于该问题，本文对模型的训练过程进行了优化，通过引入混合精度训练、对字形结构信息进行预处理和优化模型缓存等方法来加快运行速度，使得优化后的模型在训练和推理速度上与上述基线方法相差不大．

表8 融合方法对比实验

数据集	模型	*P*/%	*R*/%	*F*1/%
CCKS2017	cat	92.65	92.36	92.51
	add	87.26	88.32	87.79
	att+cat	85.83	88.27	87.05
	att+add	86.38	87.89	87.13
CCKS2019	cat	85.61	83.28	84.36
	add	85.01	83.17	84.09
	att+cat	85.03	82.71	83.87
	att+add	83.77	83.89	83.83
CMeEE	cat	67.76	68.88	68.32
	add	67.67	68.16	67.73
	att+cat	65.55	68.08	67.31
	att+add	67.22	67.00	67.11

6.5 消融研究

为验证本文所提出的方法中主要组件的有效性，如图4所示，设计了消融实验．在相同的实验环境下(RTX A4000，16 GB)，对比了完整模型(*G*)、去掉字形特征向量的模型(*-g*)、去掉负样本学习器的模型(*-n*)、去掉负样本学习器和字形特征向量的模型(*-g-n*)在CCKS2017、CCKS2019和CMeEE三个数据集上的效果．

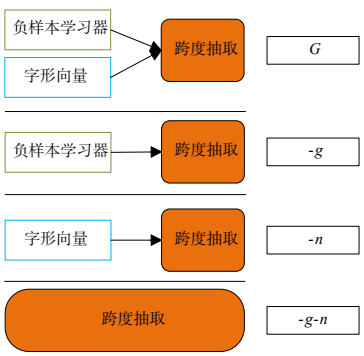


图4 消融实验对比模型

从表9中可以看出：(1)去掉字形特征向量后(-g)，模型的性能在CCKS2017、CCKS2019和CMeEE三个数据集上分别下降4.24%、0.63%和0.39%的*F*1值，说明融合字形特征向量能够有效地提升模型性能；(2)去掉负样本学习器后(*-n*)，模型分别下降4.6%、0.88%和0.79%的*F*1值，相比于去掉字形特征向量带来的模型性能下降更明显，并且在精确率*P*和召回率*R*两个指标上都出现了明显的下降，说明引入负样本学习器一方面能够减少未标记实体对模型的误导，提升模型的精确率*P*，另一方面也能够发现更多的潜在实体，提升模

型的召回率R,能够全面提升模型的性能;(3)去掉字形特征向量和负样本学习器后(*-g-n*),模型相比于仅去掉字形特征向量(*-g*)和仅去掉负样本学习器(*-n*)都出现了性能下降,说明融合字形特征向量和负样本学习器两个部分能够共同提升模型在中文医学NER任务上的性能.

表9　消融实验

数据集	模型	*P*/%	*R*/%	*F*1/%
CCKS2017	*G*	92.65	92.36	92.51
	-g	86.79	89.75	88.27
	-n	86.66	89.16	87.91
	-g-n	85.80	89.06	87.43
CCKS2019	*G*	85.61	83.28	84.36
	-g	84.08	83.39	83.73
	-n	84.69	82.43	83.48
	-g-n	82.74	82.38	82.56
CMeEE	*G*	67.76	68.88	68.32
	-g	67.68	68.17	67.93
	-n	67.22	67.84	67.53
	-g-n	66.37	67.29	66.83

7　总结

本文研究了汉字的字形结构对于医学实体的有效性以及如何利用负样本中的未标注实体.通过对医学数据集进行分析,发现医学实体在字形上特点,并证明融入字形结构特征对于解决预训练模型无法更多的包含医学领域上的动态语义这一问题是有效的.同时结合医学实体的特点设计了负样本学习器,使得模型能从负样本中学习到额外的实体信息.通过与经典的基线方法进行比较,证明了融合字形特征和引入负样本学习器对于提升模型在中文医学NER任务上的性能是有效的.未来,将考虑如何在该模型的基础上更有效地融合汉字的词信息以及继续改进负样本学习器来提升模型在中文医学NER任务上的性能.

参考文献

[1] 李冬梅, 罗斯斯, 张小平, 等. 命名实体识别方法研究综述[J]. 计算机科学与探索, 2022, 16(9): 1954-1968.
LI D M, LUO S S, ZHANG X P, et al. Review on named entity recognition[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(9): 1954-1968. (in Chinese)

[2] 杨锦锋, 于秋滨, 关毅, 等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报, 2014, 40(8): 1537-1562.
YANG J F, YU Q B, GUAN Y, et al. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction[J]. Acta Automatica Sinica, 2014, 40(8): 1537-1562. (in Chinese)

[3] ZHANG Z Y, HAN X, LIU Z Y, et al. ERNIE: Enhanced language representation with informative entities[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 1441-1451.

[4] 胡婕, 胡燕, 刘梦赤, 等. 基于知识库实体增强BERT模型的中文命名实体识别[J]. 计算机应用, 2022, 42(9): 2680-2685.
HU J, HU Y, LIU M C, et al. Chinese named entity recognition based on knowledge base entity enhanced BERT model[J]. Journal of Computer Applications, 2022, 42(9): 2680-2685. (in Chinese)

[5] 殷章志, 李欣子, 黄德根, 等. 融合字词模型的中文命名实体识别研究[J]. 中文信息学报, 2019, 33(11): 95-100, 106.
YIN Z Z, LI X Z, HUANG D G, et al. Chinese named entity recognition ensembled with character[J]. Journal of Chinese Information Processing, 2019, 33(11): 95-100, 106. (in Chinese)

[6] LI Y M, LIU L M, SHI S M. Empirical analysis of unlabeled entity problem in named entity recognition[EB/OL]. [2020]. http://arxiv.org/abs/2012.05426.pdf.

[7] LI Y M, LIU L M, SHI S M. Rethinking negative sampling for handling missing entity annotations[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 7188-7197.

[8] LING Y, HASAN S A, FARRI O, et al. A domain knowledge-enhanced LSTM-CRF model for disease named entity recognition[J]. AMIA Joint Summits on Translational Science, 2019, 2019: 761-770.

[9] LI Y, DU G D, XIANG Y, et al. Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge[J]. Journal of Biomedical Informatics, 2020, 106: 103435.

[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 Processing and Chinese Computing. Cham: Springer, 2016: 239-250.

[11] 崔少国, 陈俊桦, 李晓虹. 融合语义及边界信息的中文电子病历命名实体识别[J]. 电子科技大学学报, 2022,

51(4): 565-571.
CUI S G, CHEN J H, LI X H. Named entity recognition for Chinese electronic medical record by fusing semantic and boundary information[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(4): 565-571. (in Chinese)

[12] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2019: 4171-4186.

[13] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. [2023]. https://www. cs. ubc. ca/~amuham01/LING530/papers/radford2018-improving.pdf.

[14] GU Y J, QU X Y, WANG Z F, et al. Delving deep into regularity: A simple but effective method for Chinese named entity recognition[C]//Findings of the Association for Computational Linguistics: NAACL 2022. Stroudsburg: Association for Computational Linguistics, 2022: 1863-1873.

[15] 吴炳潮, 邓成龙, 关贝, 等. 动态迁移实体块信息的跨领域中文实体识别模型[J]. 软件学报, 2022, 33(10): 3776-3792.
WU B C, DENG C L, GUAN B, et al. Dynamically transfer entity span information for cross-domain Chinese named entity recognition[J]. Journal of Software, 2022, 33(10): 3776-3792. (in Chinese)

[16] PENG M L, XING X Y, ZHANG Q, et al. Distantly supervised named entity recognition using Positive-unlabeled learning[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 2409-2419.

[17] LI X N, YAN H, QIU X P, et al. FLAT: Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 6836-6842.

[18] 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: Association for Computational Linguistics, 2021: 1529-1539.

[19] LI X Y, FENG J R, MENG Y X, et al. A unified MRC framework for named entity recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 5849-5859.

[20] YANG P, CONG X, SUN Z Y, et al. Enhanced language representation with label knowledge for span extraction[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 4623-4635.

作者简介

孟伟伦 男,1998 年 4 月出生于河北省衡水市. 现为燕山大学信息科学与工程学院博士,主要研究方向为自然语言处理.

郭景峰 男,1962 年 2 月出生于黑龙江省哈尔滨市. 现为燕山大学信息科学与工程学院计算机系教授、博士生导师,在国内外发表学术论文 150 余篇.
E-mail: jfguo@ysu.edu.cn

邢珂萱 女,1998 年 12 月出生于黑龙江省大庆市. 现为燕山大学信息科学与工程学院研究生,主要研究方向为自然语言处理.

魏 宁 男,1995 年 1 月出生于山东省枣庄市. 现为燕山大学信息科学与工程学院博士,主要研究方向为推荐系统以及数据挖掘.

王巧梭 女,1965 年 5 月出生于河北省石家庄市. 现为河北建材职业技术学院高级工程师、高级实验师,主要研究方向计算数据分析与基础应用.

刘 滨 男,1975 年 11 月出生于河北省石家庄市. 现为河北科技大学大数据与社会计算研究中心教授、硕士生导师,在国内外发表学术论文 100 余篇.