计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(9):38−47 [doi: 10.15888/j.cnki.csa.009605]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合多特征的骨签释文实体识别①

石雨梦[1], 王慧琴[1], 王　展[2], 刘　瑞[3], 王　可[1]

[1](西安建筑科技大学 信息与控制工程学院, 西安 710055)
[2](陕西省文物保护研究院, 西安 710075)
[3](中国社会科学院 考古研究所, 北京 100101)
通信作者: 王慧琴, E-mail: hqwang@xauat.edu.cn

摘　要: 构建适用于汉长安城骨签释文的命名实体识别模型, 用来解决由于汉长安城骨签释文关键内容缺失, 而导致无法对部分骨签释文进行分类的问题. 本文将汉长安城骨签释文原始文本作为数据集, 采用 BIOE (begin, inside, outside, end) 标注方法对释文实体进行数据标注, 并提出融合字结构特征、字词结构特征的多特征融合网络模型 (multi-feature fusion network, MFFN). 该模型不仅考虑了单个字符的结构特征, 还融合了字与词的结构特征, 以增强模型对骨签释文的理解能力. 实验结果表明, MFFN 模型能够更好地识别汉长安城骨签释文的命名实体, 实现骨签释文分类, 优于现有 NER 模型, 为历史学家和研究人员提供更加丰富和准确的数据支持.

关键词: 骨签; 实体识别; BIOE 标注方法; 多特征融合; 释文分类

引用格式: 石雨梦,王慧琴,王展,刘瑞,王可.融合多特征的骨签释文实体识别.计算机系统应用,2024,33(9):38–47. http://www.c-s-a.org.cn/1003-3254/9605.html

Entity Recognition for Interpretation of Bone-sign Integrated with Multiple Features

SHI Yu-Meng[1], WANG Hui-Qin[1], WANG Zhan[2], LIU Rui[3], WANG Ke[1]

[1](College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)
[2](Shaanxi Institute for the Preservation of Cultural Heritage, Xi'an 710075, China)
[3](Institute of Archaeology, Chinese Academy of Social Sciences, Beijing 100101, China)

Abstract: This study constructs a named entity recognition (NER) model suitable for the bone-sign interpretations of Han Chang'an City to solve the problem of the inability to classify some bone-sign interpretations due to the lack of key content. The original text of the bone-sign interpretations of Han Chang'an City is used as the dataset, and the begin, inside, outside, end (BIOE) annotation method is utilized to annotate the bone-sign interpretation entities. A multi-feature fusion network (MFFN) model is proposed, which not only considers the structural features of individual characters but also integrates the structural features of character-word combinations to enhance the model's comprehension of the bone-sign interpretations. The experimental results demonstrate that the MFFN model can better identify the named entities of the bone-sign interpretations of Han Chang'an City and classify the bone-sign interpretations, outperforming existing NER models. This model provides historians and researchers with richer and more precise data support.

Key words: bone-sign; entity recognition; BIOE annotation method; multiple features fusion; classification of interpretation

汉长安城骨签出土于未央宫遗址西北部[1], 呈矩形或长方形, 长约 10–20 cm, 宽约 2–5 cm. 骨签释文, 即刻在骨签上的文字. 出土骨签共 64 305 枚, 其中刻字骨签约 57 000 枚 (如图 1), 无字骨签共近万枚, 在刻字骨

① 基金项目: 国家社科基金冷门绝学研究专项 (20VJXT001)
收稿时间: 2024-03-05; 修改时间: 2024-04-03; 采用时间: 2024-04-10; csa 在线出版时间: 2024-07-26
CNKI 网络首发时间: 2024-07-29

签中, 骨签约有 14 000 枚, 刻字骨签上刻有官职和职位, 例如, 刺史、郎中、侍卫等, 这些信息可反映出当时社会的行政组织机构以及人们的社会地位. 刻字骨签是中国考古学的重要文献, 为考古学家研究汉代历史提供了丰富的研究资料[2]. 由于骨签在地下埋藏 2 000 多年, 骨签表面有不少附着物, 尤其是一些坚硬的碳酸钙几乎与骨签连为一体, 因此必须在前期对骨签正面的“磨光面”的碳酸钙进行除锈, 再加上骨签年代久远, 并且工匠在刻字过程中, 存在“同一骨签多次刻字”“刻字未完成”“补字改字”“减字”等情况, 导致骨签释文部分内容缺失或模糊, 而部分缺失内容对骨签分类有着关键作用. 因此, 人工进行骨签释文分类时, 由于多数骨签释文内容不完整, 领域人员有时不易区分粒度细微语义相近的标签, 使得部分标签的标注结果存在冲突和遗漏, 增加了相关考古工作人员 (领域专家) 的工作量. 随着自然语言处理的发展, 应用深度学习技术对骨签释文进行实体识别、实现骨签释文自动分类成为可能.

图 1 刻字骨签

1 相关工作

1.1 古汉语命名实体识别

古汉语命名实体识别 (named entity recognition, NER)[3]是一项具有挑战性的研究领域, 能够识别和分类古汉语文本中的特定实体, 如人名、地名、时间、事件等, 减轻了研究人员在手动标注和分析工作中的负担[4]. 尽管古汉语 NER 相较于现代汉语 NER 尚处于发展初期, 但其研究和应用已取得了显著的成果. 目前, 深度学习、注意力机制和迁移学习等方法[5]已成为 NER 技术发展的主流趋势. 随着深度学习的兴起, 尤其是双向长短时记忆网络 (bidirectional long short-term memory, BiLSTM)、BERT (bidirectional encoder representations from Transformers)[6]等深度学习模型在中文 NER 问题上表现出色, 并逐渐成为研究主流. 在 NER 任务中, 研究者通常使用 $F1$ 值 ($F1$ score) 来评价模型的整体性能. $F1$ 值是精确率 (precision) 和召回率 (recall) 的调和平均数, 其中精确率是指正确识别的正类样本数占所有被识别为正类的样本数的比例, 而召回率是指正确识别的正类样本数占所有实际正类样本数的比例. $F1$ 值是一个综合考虑精确率和召回率的性能评估指标, $F1$ 值越高, 表示模型性能越稳定. 刘宇瀚等[7]结合汉字字形特征、迭代学习以及 BiLSTM 和条件随机场 (conditional random field, CRF) 的神经网络模型, 实现了与最好的基线模型相比 1.52% 的 $F1$ 指标提升. 王子牛等[8]利用 BERT 模型对人民日报语料进行实体识别, 达到了 94.86% 的 $F1$ 值. 刘新亮等[9]针对生鲜蛋供应链中的特定领域实体, 提出了 BERT 与 CRF 结合的模型, 实现了 91.01% 的 $F1$ 指标. 朱锁玲等[10]采用规则与统计相结合的方法进行 NER 任务, $F1$ 值为 71.83%. 周好等[11]采用 CRF 模型实现了古籍引书上下文的自动识别. 李成名[12]将 LSTM-CRF 模型用于《左传》中的地名和人名的自动识别, $F1$ 值分别达到 82.79% 和 82.49%. 徐晨飞等[13]基于 Bi-RNN、Bi-LSTM、Bi-LSTM-CRF 和 BERT 这 4 种深度学习模型, 对《方志物产》进行了多类实体自动识别, $F1$ 值最高达到 89.70%, 进一步证明了深度学习模型在处理古汉语 NER 任务时的高效性. 2023 年, 邓宇扬等[14]在汉藏双语 NER 的研究中, 利用预训练模型优化了 $F1$ 值, 有效识别了藏族节日文本中的实体. 武帅等[15]探讨了结合句法特征的 BERT-BiLSTM-MHA-CRF 模型, 该模型在细粒度古籍实体识别任务中表现优异, 实验结果在不同数据集上展现了良好的 $F1$ 值. Tang 等[16]利用 BERT 模型结合上下文信息提取动态词向量, 并通过 BiLSTM 模块进一步训练, 在 MASR 的 $F1$ 值提高了 0.55%. Zhang 等[17]提出了多粒度 BERT 适配器和高效全局指针, 针对传统注意力机制中在捕捉边界信息方面的局限性, 通过增强模型对嵌套实体结构的识别能力, 显著提高了识别的准确度.

1.2 汉长安城骨签释文内容的研究

汉长安城骨签类别多样, 根据释文内容可分为 5 大类别, 分别是工官类、编号类、计量单位类、弓弩名称类, 以及中央官署类. 其中的骨签数量高达 14 076 枚, 约占刻字骨签总数的 1/3. 在工官类骨签中, 骨签分

为“河南工官”“南陽工官”“穎川工官”这 3 大类. 由于骨签释文关键文字的缺失, 导致无法对其进行准确的实体识别, 在对其进行分类的过程中, 遇到的问题主要分为以下 3 种.

(1) 缺失文字相对位置确定. 如图 2 所示, 部分骨签释文会出现“□南工官”“河□工官”“南□工官”“□川工官”的文字缺失情况, 这类骨签虽缺失文字, 但文字相对位置确定, 依据人工可判定其分别为“河南工官”“河南工官”“南陽工官”“穎川工官”的骨签.

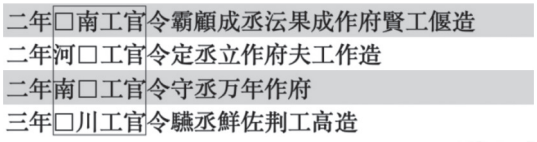


图 2 释文文本缺失情况 1

(2) 缺失文字相对位置不确定, 关键字确定. 如图 3 所示, “河工官”“陽工官”“穎工官”“川工官”, 这类骨签虽缺失文字, 但依据保存完好的文字仍能判定其分别为“河南工官”“南陽工官”“穎川工官”“穎川工官”的骨签.

(3) 缺失文字相对位置不确定, 关键字仍不确定. 如图 4 所示, “南工官”这类骨签由于关键字“河”或“陽”的缺失, 根据保存的文本无法确认其为“河南工官”类或“南陽工官”类.

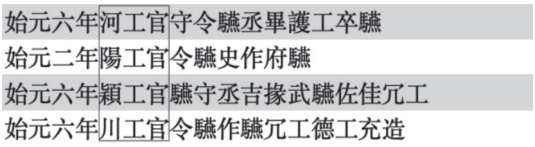


图 3 释文文本缺失情况 2

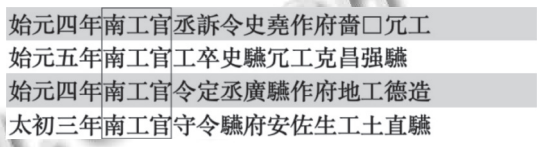


图 4 释文文本缺失情况 3

基于上述分析, 本研究提出一种融合字结构特征与字词结构特征的多特征融合网络模型 (multi-feature fusion network, MFFN), 对缺失关键文字而导致无法进行人工分类的“南工官”类骨签释文进行实体识别, 实现骨签释文的分类, 并设计两组对比实验进行分析, 进一步来验证本文提出模型的有效性, 实验分析维度如图 5 所示.

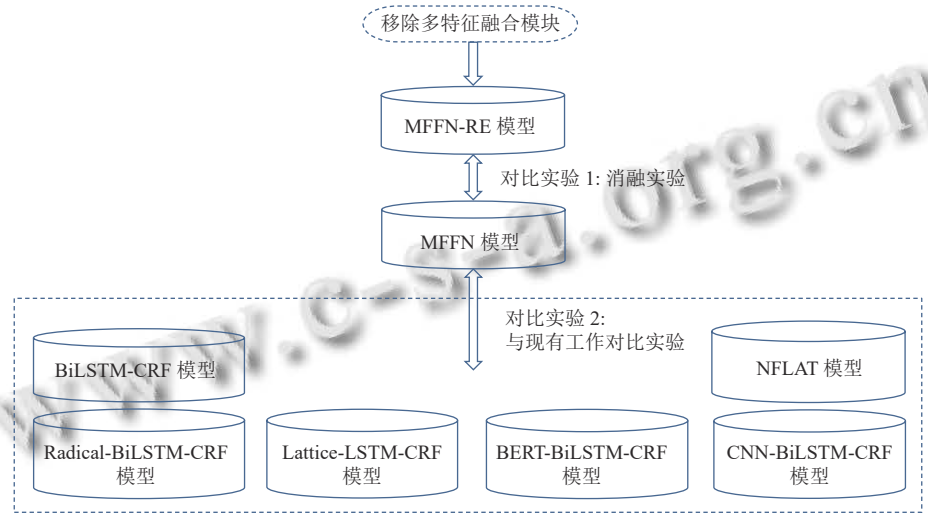


图 5 实验分析维度

1.3 骨签释文数据集的构建

1.3.1 数据采集

本实验数据集的来源是汉长安城骨签释文, 由于实验的最终目的是实现关键文字缺失的骨签释文的实体识别, 而仅通过人工很难实现对骨签释文的正确实体识别, 因此选取关键字保存相对完整的骨签释文作为本次实验的数据集, 命名实体的类别为两大类, 分别是河南工官、南陽工官, 共 10 040 条数据. 通过 Python 对骨签释文进行批量读取并进行人工标注, 将其存储为 JSON 文件. 每个 JSON 文件的每一个 JSON 数组代表一条单独的数据, 每条数据包括一个原始句子以及句子中命名实体的标签, 具体的数据格式如图 6 所示.

```
"text": "始元年河南工官守丞作府斋夫佐郢",
"label": {
  "henan": {
    "河南工官": [[3, 6]]
  }
}
```


图 6 骨签释文标注数据格式

1.3.2 数据预处理

原始的骨签释文文本数据, 包含了许多无用信息, 如标点符号和字符等, 给后续实体识别工作带来了较大干扰, 故在进行骨签释文的实体识别工作前对释文进行预处理至关重要. 对骨签释文原始语料库进行数据清洗来获取比较规范的数据集, 数据清洗主要包括以下两部分.

第 1 部分: 对短文本进行清洗过滤, 删除“/”“…”等对骨签释文文本分析无意义的符号和语句, 以此来减少数据噪声.

第 2 部分: 对短文本进行去停用词处理, 删除诸如“驨”等意义不大的词汇, 减少文本的冗余度.

1.3.3 实体标注

我国古汉语 NER 技术不断发展与进步, 但由于不同的研究领域所对应的文本数据都有各自的文本特征, 因此必须针对领域特征, 设计一种适应数据特点的文本标注方法[18]. 通过对汉长安城骨签释文文本的实体结构进行分析, 决定采用 BIOE 标注体系 (B-begin, I-inside, O-outside, E-end) 与实体类别相结合的方式, 根据目标词在上下文中的含义进行标记, 并且实体标注集用 S 表示, 具体为 S={B, I, E}, 数据集标注样例格式如表 1 所示.

表 1 数据集标注样例

原始文本	标签	标签解释
二	O	非实体
年	O	非实体
河	B-henan	河南工官开头
南	I-henan	河南工官内部
工	I-henan	河南工官内部
官	E-henan	河南工官结尾
守	O	非实体
丞	O	非实体

2 骨签释文实体识别模型的构建

2.1 汉长安城骨签释文文本结构分析

汉长安城骨签释文在结构和句式方面与现代文本或普通汉语文本存在显著差异. 通过深入分析骨签释文文本, 可以发现存在以下特征.

特征 1: 骨签释文结构复杂, 并存在多数单个汉字作为独立单位的情况. 因此为了确保骨签释文中的单个汉字能够在特征表示中保持其独立性, 同时避免由于分词错误引入的不良影响, 决定采用单个汉字作为模型的输入向量 (即字符嵌入), 以便于捕获骨签释文文本的形态学和词法信息.

特征 2: 骨签释文文本信息密度较高, 并且释文的下文信息对工官实体的识别更为重要, 而单向的 LSTM 网络只会捕获到释文文本左边的输入信息, 无法将释文文本右边的输入信息加入模型中考虑. 为了充分利用骨签释文文本前后文, 特别是后文的语境信息, 选择 BiLSTM 网络模型进行上下文语义特征提取. 通过使用 BiLSTM 网络结构, 利用前向和后向两个方向的 LSTM 网络提取骨签释文文本信息的特征, 实现对骨签释文文本上下文信息的关联.

特征 3: 骨签释文文本中存在同一词可能有多个可能的标签. 例如在“二年南工官謝丞驨定作府輔工楚造”骨签释文中, “工”字既表示为实体中的“I”标签, 也表示为与实体无关的“O”标签. 条件随机场 CRF 可以帮助解决骨签释文标签歧义问题, 通过考虑上下文信息来确定最合适的标签, 对标签之间设置一些合法的约束性条件.

2.2 模型总框架设计

本文提出融合字结构特征和字词结构特征的多特征融合网络 (MFFN), 模型总体框架如图 7 所示.

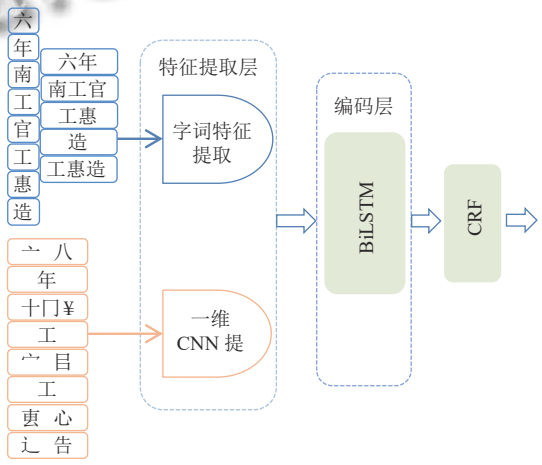


图 7 MFFN 模型总体框架

模型主要由 3 个部分组成: 特征提取层、BiLSTM 编码器和 CRF 解码器. 特征提取层用于提取骨签释文

的文本特征, BiLSTM 编码器从输入层的向量中学习骨签释文数据中的实体上下文特征, CRF 解码器将 BiLSTM 层输出的特征向量转换为有序的标签序列, 得到骨签释文实体.

2.2.1 字结构特征提取

表 2 展示了骨签释文结构数据示例. 该数据包含了 235 个汉字的字结构拆分数据, 以文本形式存储.

表 2 骨签释文结构数据示例

汉字	结构拆分	
賜	贝	易
府	广	付
佗	亻	它
通	辶	甬
狀	丬	犬
謁	言	曷

为了得到骨签释文文本部件的向量化表示, 将 235 个文本的字结构进行拆分, 并将其做成一个字典, 该字典包含每一个汉字的部件向量表示.

将句子作为模型的输入, 通过字典查找句子中每个汉字的部件所对应的向量, 以获得句子级的字结构向量. 将其输入到一维卷积神经网络中, 经过卷积后使用最大池化和全连接层来获得这个句子中每个汉字的字结构特征.

2.2.2 字词结构特征提取

目前针对古汉语自然语言处理量字、词向量库尚未建立, 因此本节采用与 FLAT 模型相同的字、词向量库. 该库包含了现代汉语和古汉语的字词, 并将它们以向量化形式存储. 图 8 给出了一个示例, 模型的输入为“六年南工官工惠造”. 将句子分解为单字集合, 即“六/年/南/工/官/工/惠/造”, 将每个字与字向量库进行匹配, 得到句子的单字嵌入. 将句子与词向量库进行匹配, 提取句子中在词典存在的词汇, 即“六年/南工官/工惠/造工惠造”, 匹配后得到句子的词嵌入.

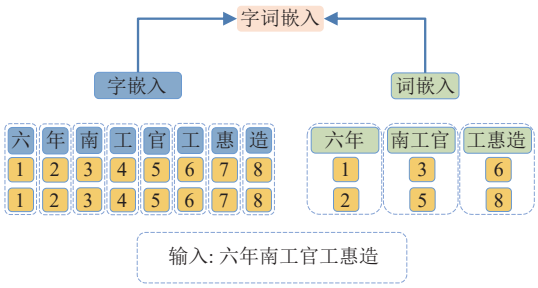


图 8 字词嵌入

2.2.3 双向长短时记忆层 (BiLSTM 编码器)

长短时记忆网络 (long short-term memory, LSTM) 能够在当前时间步保留前一时间步的信息. LSTM 通过遗忘门 (f_t)、记忆门 (i_t)、输出门 (o_t) 和细胞状态 (C_t) 之间的相互作用, 有效提取骨签释文数据的特征. 其中W是遗忘门的权重矩阵, b是偏置项, σ是激活函数. 算法流程如下.

(1) 计算遗忘门 (f_t). 遗忘门决定从上一时间步的细胞状态中保留哪些信息. 通过将前一时间步的细胞状态h_{t-1}和当前时间步的输入向量X_t相结合, 与遗忘门的权重矩阵W_f相乘, 再结合偏置项, 通过激活函数σ得到输出f_t, 即:

$$f_t = \sigma(W_f \times [h_{t-1}, X_t] + b_f) \tag{1}$$

(2) 计算记忆门 (i_t). 记忆门决定在当前时间步哪些新的信息将被存储到细胞状态中. 输入为前一时间步的细胞状态h_{t-1}、当前输入向量为X_t, 输出为记忆门的值i_t与临时细胞状态$\tilde{C}_t$, 通过记忆门决定是否将当前时间步的输入信息添加到记忆状态中, 即:

$$i_t = \sigma(W_i \times [h_{t-1}, X_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_c \times [h_{t-1}, X_t] + b_c) \tag{3}$$

(3) 计算当前细胞状态 (C_t). 当前时间步的细胞状态C_t是由上一时间步的细胞状态C_{t-1}通过遗忘门f_t, 最后通过记忆门i_t和临时细胞状态$\tilde{C}_t$而得到的, 即:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{4}$$

(4) 计算输出门 (o_t) 和当前隐层状态 (h_t). 输出门o_t和当前隐层状态h_t由前一时间步的隐层状态h_{t-1}、当前输入X_t以及当前细胞状态C_t共同决定, 即:

$$o_t = \sigma(W_o [h_{t-1}, X_t] + b_o) \tag{5}$$

$$h_t = o_t \times \tanh C_t \tag{6}$$

最终采用与句子长度一致的隐层状态序列来编码输入句子, 以此捕捉句子的上下文信息.

骨签释文实体的识别不仅与前文的信息有关, 后文的信息更为重要, 即在模型的训练中, 需要访问过去以及未来的输入特征. LSTM 网络通过长序列记忆功能处理骨签释文数据, 但其单向性限制了模型对后文信息的充分利用. BiLSTM 通过结合正向与反向传播路径, 实现了对骨签释文数据全面的特征学习. 在 BiLSTM 结构中, 文本序列进行正向遍历和反向遍历,

并将两个方向的隐层状态进行拼接, 以捕捉每个时间点的全面上下文信息. 这种方法有效地解决了传统 LSTM 在处理长文本时忽略后续信息的问题. 如图 9 所示, 经过嵌入层后, 向量化的文本进入 BiLSTM 进行上下文语义特征提取.

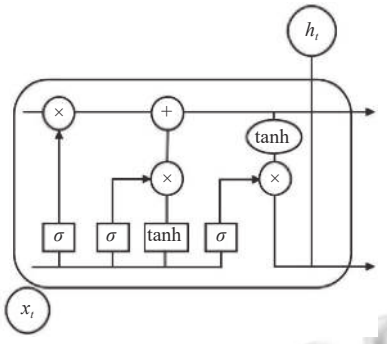


图 9 BiLSTM 模型框架

2.2.4 条件随机场 CRF 层

CRF 模型是解决序列标注问题的经典模型. 传统 NER 任务中, 利用 BiLSTM 的输出进行归一化处理, 以预测每个时间点上最可能的标签序列. 然而, 这种独立标签预测方法可能会产生非法的标签序列, 例如: 骨签释文实体的头部本应标记为“B”标签 (表示为实体的开始标签), 而被错误地标记为“I”标签 (表示为实体的内部标签). 为了确保标签序列的合法性, 同时使模型能够学习到标签间的依赖关系, 将 BiLSTM 的隐层输出特征输入到 CRF 层. CRF 层通过引入状态转移矩阵, 对标签之间的转移进行建模, 从而在全局范围内寻找最优的标签序列, 确保序列标注的一致性和准确性. 这种方法有效地整合了“BIOE”标注体系中标签间的约束, 提升了模型在序列标注任务中的性能.

对于任一个序列$X=(x_1,x_2,\cdots,x_k)$, BiLSTM 解码器输出的矩阵为P, P的大小为$m\times n$, 其中m为词的个数, 为n标签个数, P_{ij}表示第i个词的第j个标签的分数. 对预测序列$Y=(y_1,y_2,\cdots,y_k)$而言, 得到它的分数函数为:

$$s(X,Y)=\sum_{i=0}^{k}A_{y_i,y_{(i+1)}}+\sum_{i=1}^{k}P_{i,y_i} \tag{7}$$

其中, A表示转移分数矩阵, A_{ij}代表标签i转移为标签j的分数, A的大小为$n+2$, 预测序列Y产生的概率为:

$$p(Y\mid X)=\frac{e^{s(X,Y)}}{\sum_{\tilde{Y}\in Y_X}s(X,\tilde{Y})} \tag{8}$$

其中, $\tilde{Y}$表示真实的标注序列, Y_X表示所有可能的标注序列. 解码后地最大分数的输出序列:

$$Y^*=\arg\max s(X,Y) \tag{9}$$

3 实验与结果分析

3.1 实验环境和参数设置

本实验基于 Torch 框架, 并且实验的所有模型均采用同一实验环境进行训练, 具体实验环境设置如表 3 所示.

表 3 实验环境设置

项目	环境
操作系统	Windows 10 x64
处理器	Inter(R) Xeno(R) Silver4210R CPU @ 2.40 GHz
内存	8 GB
Python版本	3.8
Torch版本	2.1.0

在深度学习模型中, 参数的选择会显著决定模型的质量, 通过实验调整模型中的各项参数值并观察测试集的实体识别效果, 最终选取模型主要的训练参数设置如表 4 所示.

表 4 参数设置

参数名称	参数值
Embedding-size	64
Hidden-size	128
Learning_rate	0.001
Dropout	0.5
Batch_size	64
Epoch	30
Optimizer	Adam

Embedding-size 参数: Embedding-size 表示嵌入向量的维度, 这个向量包含词汇或字符的语义信息, Embedding-size 的选择对于模型性能和计算效率有着重要的影响. 由于本实验选用的是基于字符嵌入的 NER 方法, 而字符嵌入处理的是字符级别的信息, 通常拥有较小的维度, 最终将 Embedding-size 设置为 64.

Hidden-size 参数: Hidden-size 决定了 BiLSTM 层隐藏状态的维度大小, 在 BiLSTM 模型中, Hidden-size 用来捕获输入序列中的长期依赖关系, 它的大小会直接影响模型的记忆能力, 再结合本实验属于字符级别的任务, 最终将 Hidden-size 设置为 128.

Learning_rate 参数: Learning_rate 表示在每次迭代中模型参数更新的幅度和步长, 对于避免 BiLSTM 模

型的梯度消失或爆炸现象至关重要. 为了缓解 BiLSTM 模型在训练过程中的遗忘问题, 将 Learning_rate 设置为 0.000 1 进行实验, 逐渐增加学习率, 提高模型的收敛性, 最终将 Learning_rate 设置为 0.001. 同时设置参数 Dropout 为 0.5, 减少隐藏层中神经元之间的相关性, 减少过拟合风险.

Batch_size 参数: Batch_size 对模型的泛化能力有显著影响, 增大 Batch_size 可以减少训练周期并增强模型的稳定性, 但同时会增加内存消耗. 本实验的数据量庞大, 内存空间足够, 因此最终将 Batch_size 设置为 64.

Epoch 参数设置: 使用早停策略 (early stopping), 即在验证集上监测模型性能, 一旦性能有下降趋势, 立刻停止训练, 基于这一策略进行实验, 最终将 Epoch 参数设置为 30, Optimizer 为 Adam 优化器.

3.2 评价指标

实体识别任务中, 只有同时正确识别实体类型与实体边界, 才能被算作成功的实体识别. 以实体“河南工官”为例, 经过人工标注后, 实体对应的标注为“B-henan”“I-henan”“I-henan”“E-henan”. 对标注结果进行评估时, 只有当评估的 4 个标签与标注集完全一致时, 才能被视为正确识别一个实体. 即评判一个实体是否被正确识别, 主要包括以下 3 个方面.

(1) 实体边界是否正确 (实体的开始位置标注为“B-”, 结束位置标注为“E-”);

(2) 实体类别是否一致 (实体类别全部为“nanyang”或“henan”);

(3) 实体内部位置是否标注正确 (实体内部标注为“I-”).

在训练和测试骨签释文数据中, 通过统计所有实体的识别结果的个数, 包括正确识别和错误识别的个数, 使用自然语言处理中常用的评估指标 $F1$ 值, 来评估模型的性能. 实验结果的分类评价指标如表 5 所示.

$F1$ 值的计算式如下:

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{10}$$

其中, P 和 R 分别代表精确率和召回率. P 和 R 的计算式为:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{11}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{12}$$

表 5 分类评价指标含义

类别	预测正类	预测反类
实际正类	*TP*	*FN*
实际反类	*FP*	*TN*

精确率表示正确预测为正类别的比例, 精确率越高, 表示模型对负类别的区分能力越强; 召回率表示正确预测为正类别的样本占所有实际正类别样本的比例, 召回率越高, 表示模型对正类别的识别能力越强; 而 $F1$ 值综合考虑了精确率和召回率, $F1$ 值越高, 表示模型性能越稳定.

3.3 实验设计与结果

为验证本文提出模型的性能, 设置了 2 组对比实验.

(1) 消融实验. 为探究本文提出模型的有效性, 对其进行消融实验.

(2) 与现有工作的对比实验. 将本文所提出的模型与现有其他工作提出的模型进行对比, 验证本文模型对汉长安城骨签释文实体识别任务的有效性.

3.3.1 消融实验

为验证 MFFN 模型对汉长安城骨签释文命名实体识别的有效性, 对其进行了消融实验, 其中 MFFN-RE 是去除多特征融合后的模型, 实验结果如表 6 所示.

表 6 消融实验结果 (%)

模型	*P*	*R*	*F*1
MFFN-RE	85.36	82.87	83.56
MFFN	92.43	91.02	91.35

实验结果表明, BiLSTM-CRF 的各项指标都低于 MFFN, 说明多特征融合模块能够更深层地提取骨签释文文本特征, 有助于提升实体的识别率.

3.3.2 与现有工作的对比实验

在上述实验的基础上, 为进一步验证本研究提出的 DA-BiLSTM-CRF 模型的性能, 将其与现有其他工作在汉长安城骨签释文命名实体识别的效果进行对比, 实验结果如表 7 所列.

以上数据结果表明, 各类实体识别模型在识别效果上基本取得了不错的成绩. 本文提出的 MFFN 模型总体的性能最优, 相比基础模型 BiLSTM-CRF 模型在 $F1$ 上提高了 9.38%, 这说明骨签释文字结构中确实存在着一定的语义, 对骨签释文实体识别起到了积极作用. 而 BERT 是在大规模文本语料库上进行预训练的, 它的预训练是基于通用文本而不是特定领域的, 模型的性能在很大程度上受到预训练模型的质量和数据覆

盖范围的影响. 目前, 尚未存在古汉语预训练模型, 因此 BERT 无法充分捕获古汉语的语言特性. 相比于将字根序列与字向量拼接组成模型输入的 Radical-BiLSTM-CRF 模型, 其 $F1$ 值提高了 8.4%.

表 7　本文模型与现有其他模型之间的性能比较 (%)

模型	P	R	$F1$
CNN-BiLSTM-CRF[19]	82.58	84.78	83.94
Radical-BiLSTM-CRF[20]	83.68	81.33	82.23
Lattice-LSTM-CRF[21]	86.44	83.83	86.87
BERT-BiLSTM-CRF[22]	79.68	79.94	80.83
BiLSTM-CRF[23]	74.96	66.15	75.97
NFLAT[24]	86.77	90.62	87.29
MFFN	92.43	91.02	91.35

此外, 对比分析了各模型前 30 个 Epoch 的 $F1$ 值. 如图 10 所示, BiLSTM-CRF 模型和 NFLAT 模型在初期的 $F1$ 值较低, 随着训练次数的增加, 才逐渐上升最终稳定; 而其他对比模型在训练初期的 $F1$ 值相对较高, 之后逐渐提升并稳定, 但无法超过 MFFN 模型.

3.3.3　实验结果验证

对骨签释文原始图像进行分析, 通过骨签释文特征将其进行分类, 再结合 MFFN 模型的实体识别结果, 通过对比模型的预测结果与实际标注, 进而验证本文提出模型的准确性.

(1) 依据骨签刻字风格分析. 如图 11, 其中 (a) 为缺失关键字的 12160 号骨签释文原始图像, 其余均为释文内容保存完整的骨签释文原始图像. 依据刻字风格可以判定 8 枚骨签为同一名工匠刻写, (b)–(h) 骨签释文都包含“河南工官”的刻字信息, 因此可判定 12160 号骨签为“河南工官”类. 这种依据刻字风格分析的方法为验证 MFFN 模型的实体识别准确性提供了有效的途径.

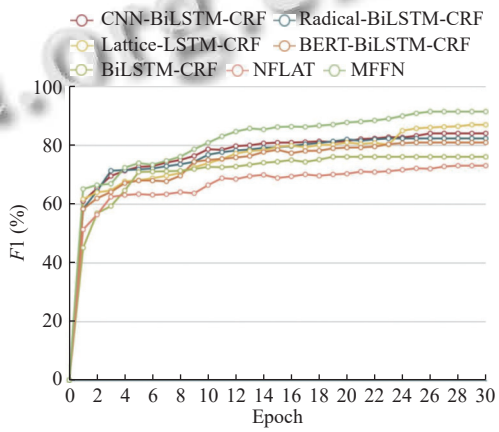
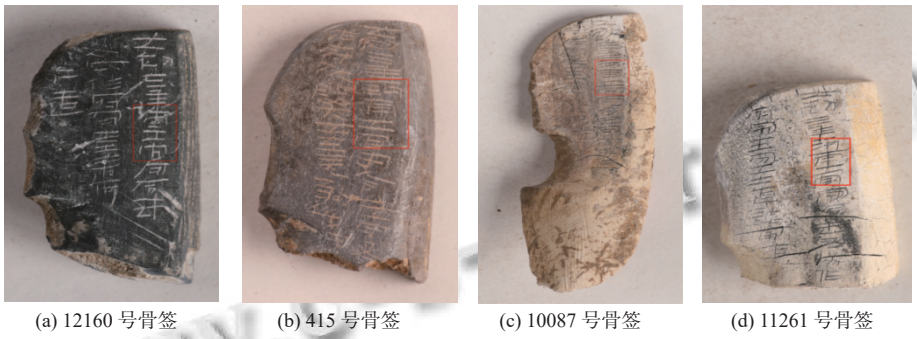
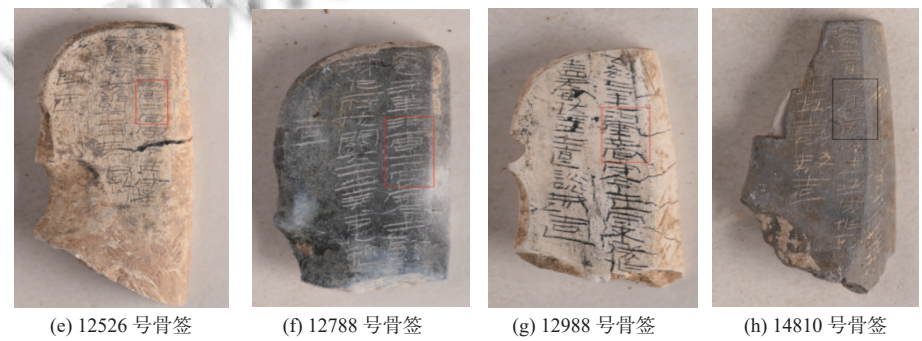


图 10　各模型 $F1$ 值更新情况



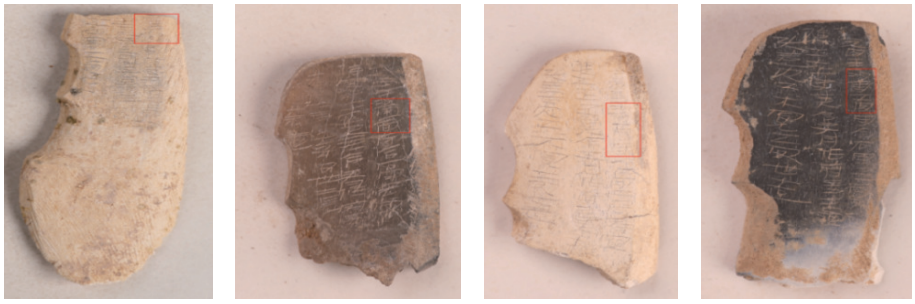
(a) 12160 号骨签　(b) 415 号骨签　(c) 10087 号骨签　(d) 11261 号骨签



(e) 12526 号骨签　(f) 12788 号骨签　(g) 12988 号骨签　(h) 14810 号骨签

图 11　骨签刻字风格

(2) 依据骨签释文语义分析. 图 12(a) 为 14217 号骨签的下半部原始图像, 由于骨签断裂, 骨签释文只保留了“南工官”的关键字, 通过分析图 12(b)–(d) 骨签的释文语义, 可以判定 12526 号骨签为“河南工官”类.



(a) 14217 号骨签 (b) 1308 号骨签 (c) 1437 号骨签 (d) 1605 号骨签

图 12 骨签刻字语义

将各模型对骨签的实体识别结果与上述分析结果进行比对, 结果如表 8 所示.

表 8 各模型比对后的准确率 (%)

模型	*F*1
CNN-BiLSTM-CRF[19]	82.55
Radical-BiLSTM-CRF[20]	84.95
Lattice-LSTM-CRF[21]	80.94
BERT-BiLSTM-CRF[22]	76.83
BiLSTM-CRF[23]	80.97
NFLAT[24]	72.11
MFFN	90.67

以上数据结果表明, MFFN 模型的实体识别结果与骨签原始图像分析结果的比对效果最优, 模型总体的准确率为 90.67%.

4 结论

本文针对汉长安城骨签释文关键内容缺失, 而导致无法对骨签进行分类的问题, 提出适用于汉长安城骨签释文的命名实体识别模型. 其中特征提取层提取骨签释文的文本特征, BiLSTM 编码器从输入层的向量中学习骨签释文命名实体的上下文特征, 最终由 CRF 解码器将 BiLSTM 层输出的特征向量转化成序列标签, 得到骨签释文命名实体, 进而实现骨签释文的分类. 实验结果表明, MFFN 模型能够更好地识别汉长安城骨签释文的命名实体, 实现骨签释文分类, 证明了 MFFN 模型的有效性. 虽然 MFFN 模型在汉长安城骨签释文命名实体识别中取得了较好的效果, 但对于骨签释文相对较短的文本, 模型识别效果明显降低, 因此在后续研究中, 考虑进一步融入骨签释文的其他语义特征、词序特征等多特征知识, 充分发挥组合模型的性能, 进而提升骨签释文实体识别效果.

参考文献

1 李毓芳. 汉长安城未央宫骨签述略. 人文杂志, 1990(2): 99–102.

2 张戈. 汉长安城骨签校注 [硕士学位论文]. 北京: 首都师范大学, 2012.

3 Li J, Sun AX, Han JL, *et al*. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50–70. [doi: 10.1109/TKDE.2020.2981314]

4 Wang Q, Zhou YM, Ruan T, *et al*. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. Journal of Biomedical Informatics, 2019, 29: 103133.

5 陈曙东, 欧阳小叶. 命名实体识别技术综述. 无线电通信技术, 2020, 46(3): 251–260. [doi: 10.3969/j.issn.1003-3114.2020.03.001]

6 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

7 刘宇瀚, 刘常健, 徐睿峰, 等. 结合字形特征与迭代学习的金融领域命名实体识别. 中文信息学报, 2020, 34(11): 74–83. [doi: 10.3969/j.issn.1003-0077.2020.11.010]

8 王子牛, 姜猛, 高建瓴, 等. 基于 BERT 的中文命名实体识别方法. 计算机科学, 2019, 46(S2): 138–142.

9 刘新亮, 张梦琪, 谷情, 等. 基于 BERT-CRF 模型的生鲜蛋供应链命名实体识别. 农业机械学报, 2021, 52(S1): 519–525. [doi: 10.6041/j.issn.1000-1298.2021.S0.066]

10 朱锁玲, 包平. 方志类古籍地名识别及系统构建. 中国图书馆学报, 2011, 37(3): 118–124.

11 周好, 王东波, 黄水清. 古籍引书上下文自动识别研究——以注疏文献为例. 情报理论与实践, 2021, 44(9): 169–175. [doi: 10.16353/j.cnki.1000-7490.2021.09.024]

12 李成名. 基于深度学习的古籍词法分析研究 [硕士学位论文]. 南京: 南京师范大学, 2018. [doi: 10.27245/d.cnki.gnjsu.2018.000211]
13 徐晨飞, 叶海影, 包平. 基于深度学习的方志物产资料实体自动识别模型构建研究. 数据分析与知识发现, 2020, 4(8): 86–97.
14 邓宇扬, 吴丹. 面向藏族传统节日的汉藏双语命名实体识别研究. 数据分析与知识发现, 2023, 7(7): 125–135.
15 武帅, 杨秀璋, 何琳, 等. 基于句法特征和 BERT-BiLSTM-MHA-CRF 的细粒度古籍实体识别研究. 数据分析与知识发现: 1–16. http://kns.cnki.net/kcms/detail/10.1478.G2.20240313.1314.004.html. [2024-04-04].
16 Tang XY, Huang Y, Xia M, *et al*. A multi-task BERT-BiLSTM-AM-CRF strategy for Chinese named entity recognition. Neural Processing Letters, 2023, 55(2): 1209–1229. [doi: 10.1007/s11063-022-10933-3]
17 Zhang L, Xia PF, Ma XX, *et al*. Enhanced Chinese named entity recognition with multi-granularity BERT adapter and efficient global pointer. Complex & Intelligent Systems, 2024. [doi: 10.1007/s40747-024-01383-6]
18 Roy A. Recent trends in named entity recognition (NER). arXiv:2101.11420, 2021.
19 Jia YZ, Xu XB. Chinese named entity recognition based on CNN-BiLSTM-CRF. Proceedings of the 9th IEEE International Conference on Software Engineering and Service Science. Beijing: IEEE, 2018. 1–4.
20 Lin CY, Xue NW, Zhao DY, *et al*. Natural Language Understanding and Intelligent Applications. Cham: Springer, 2016. 239–250.
21 Zhang Y, Yang J. Chinese NER using lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 1554–1564.
22 谢腾, 杨俊安, 刘辉. 基于 BERT-BiLSTM-CRF 模型的中文实体识别. 计算机系统应用, 2020, 29(7): 48–55. [doi: 10.15888/j.cnki.csa.007525]
23 Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 260–270.
24 Wu S, Song XN, Feng ZH, *et al*. NFLAT: Non-flat-lattice Transformer for Chinese named entity recognition. arXiv: 2205.05832, 2022.

(校对责编: 张重毅)