计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(7):195−207 [doi: 10.15888/j.cnki.csa.009886] [CSTR: 32024.14.csa.009886]

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

面向语义缺失的骨签释文分类算法①

窦相宜[1], 王慧琴[1], 王　可[1], 刘　瑞[2], 王　展[3]

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

摘　要: 陕西省西安市汉长安城遗址出土的骨签为西汉历史的研究工作提供了丰富资料, 受长期埋藏和人为开采影响, 大量骨签存在断裂现象, 造成语义信息缺失, 影响骨签分类归置效率. 为提高骨签分类归置效率, 本文提出了一种面向语义缺失的 EWRCA 骨签释文分类模型. 该模型利用 ERNIE 的 8 层编码器捕获文本的深层语义信息, 学习断裂和不完整的骨签释文信息; 通过融合 ERNIE 多层编码器的输出与 Word2Vec 生成的词向量, 提高对骨签释文独有词汇的理解能力; 将文本向量融合模块与 TextRCNN-MHAtt 模型结合, 有效捕获文本的上下文依赖, 增强文本的语义表示能力, 提升分类准确性; 引入融合注意力机制提高模型在处理骨签释文时的准确性. 实验结果表明, 该模型对汉长安城骨签释文的分类精度和准确率达到 95.62%、95.2%, 能够有效提高骨签释文的分类精度.

关键词: 骨签释文; 释文分类; ERNIE; Word2Vec; TextRCNN-MHAtt

引用格式: 窦相宜,王慧琴,王可,刘瑞,王展.面向语义缺失的骨签释文分类算法.计算机系统应用,2025,34(7):195–207. http://www.c-s-a.org.cn/1003-3254/9886.html

Classification Algorithm for Bone Stick Interpretation with Semantic Deficiency

DOU Xiang-Yi[1], WANG Hui-Qin[1], WANG Ke[1], LIU Rui[2], WANG Zhan[3]

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

Abstract: The bone sticks excavated from the Han Chang'an City site in Xi'an City, Shanxi Province, have provided valuable information for research into the history of the Western Han Dynasty. Due to the long-term burial and artificial excavation, many of the bone sticks are fractured, resulting in the loss of semantic information, which affects the efficiency of bone stick classification. To improve the efficiency of bone stick classification and imputation, a semantically deficient EWRCA bone stick interpretation classification model is proposed. The model utilizes ERNIE's 8-layer encoder to capture the deep semantic information of the text, learning from the fractured and incomplete bone stick interpretation data. By integrating the outputs of ERNIE's multi-layer encoder with the word vectors generated by Word2Vec, the model's ability to understand the unique vocabulary of bone stick interpretations is enhanced. The combination of the text vector fusion module with the TextRCNN-MHAtt model effectively captures contextual dependencies, strengthening the semantic representation of the text and improving classification accuracy. The fusion attention mechanism is introduced to enhance the accuracy and robustness of the model when processing bone stick interpretations. Experimental results demonstrate that the model achieves classification precision and accuracy of 95.62% and 95.2% for the bone stick interpretations from Han Chang'an City, significantly improving classification accuracy.

Key words: bone stick interpretation; interpretation classification; ERNIE; Word2Vec; TextRCNN-MHAtt

① 基金项目: 国家社科基金冷门绝学研究专项 (20VJXT001)
收稿时间: 2024-11-26; 修改时间: 2024-12-17, 2025-01-15; 采用时间: 2025-01-21; csa 在线出版时间: 2025-05-29
CNKI 网络首发时间: 2025-05-29

汉长安城骨签出土于未央宫遗址西北部, 将动物骨骼加工成长条形骨片, 分为正面和背面, 大多数正面刻有文字, 故称为“骨签”[1]. 古西汉时期的官方文献承载了大量关于行政管理、经济活动和社会生活的详细记录. 它们详细记载了车马、衣物、器械、兵器等多方面的登记情况, 反映了西汉时期手工业的水平和行业分布, 以及不同年代的进贡物品在数量、品种和质量上的变化. 这些信息不仅对研究汉代的经济发展具有重要的参考价值, 还可以深入了解当时的社会结构和政治情况. 但由于长时间的埋藏以及人工挖掘的影响, 使得许多骨签出现了断裂现象, 导致骨签释文内容部分缺失, 为骨签的分类带来了挑战. 传统的碎片分类方法主要依赖于考古学家的经验和直觉, 这种方法不仅耗时长, 且容易出错. 由于骨签碎片出土约 60 000 多片, 其中有文字的骨签有 57 000 余片[2], 故利用骨签释文进行骨签分类极其必要. 骨签释文, 即篆刻在骨签表面的文字 (如图 1), 其记载内容覆盖整个西汉时期, 对于研究西汉历史具有深远意义. 对骨签释文进行自动分类, 不仅能将骨签碎片更好地分类归置, 实现西汉文化遗产的保护, 还可以促进相关历史文献的整理与研究, 帮助学者们更准确地理解西汉文化和社会生活.

图 1 刻字骨签

1 相关工作

1.1 文本分类相关工作

文本分类旨在将给定文本自动归类到预定义类别. 随着深度学习和自然语言处理技术的发展, 文本分类得到广泛应用, 如垃圾邮件过滤、情感分析、话题检测等. 早期文本分类方法主要依赖于传统的机器学习算法, 如支持向量机 (SVM) 和朴素贝叶斯 (naive Bayes), 通常结合词袋模型 (bag-of-words) 或 TF-IDF 等特征表示. 虽然在处理低维特征方面表现良好, 但在面对复杂语义和高维数据时, 需要在高维特征空间进行操作, 计算成本较高. 近年来, 基于深度学习的方法取得显著进展, 许多学者开始将深度学习应用于文本分类任务. 尤其是卷积神经网络 (CNN)、循环神经网络 (RNN) 和基于注意力机制的 Transformer 模型, 显著提高了文本分类的准确性和鲁棒性. 但这些模型依然面临语义理解不够精确、对上下文关系建模能力不足等问题.

2019 年 Devlin 等人[3]提出的 BERT (bidirectional encoder representations from Transformers) 标志着一个重大突破, 引领了自然语言处理领域的新方向. BERT 基于双向 Transformer 结构, 通过在大规模语料上进行预训练, 能够捕获句子中深层次的语义关系和上下文依赖, 显著提升了对复杂语义的理解能力. 之后, 不同研究者将 BERT 及 BERT 变体应用于各种领域的文本处理任务, 效果得到显著提升. 这些 BERT 变体通过在更具针对性的领域语料库上进行预训练, 增强了模型对专业领域的理解和分类效果. Yu 等人[4]将 BERT 模型应用于政策文本分类, 通过提取政策书中句子级的特征向量, 学习文本的关键特征, 显著提升了政策文本数据集分类任务的准确性. Cui 等人[5]通过结合 BERT 预训练模型与卷积神经网络 (CNN) 提取文本特征, 有效利用了 BERT 的深层语义理解和 CNN 的局部特征提取能力, 对中文文本实现了有效分类. 此外, 郝婷等人[6]提出了一种结合 BERT 和 Bi-LSTM 模型, 利用 BERT 的深层语义分析和 Bi-LSTM 解决句子中长距离依赖问题, 提高了新闻短文本的分类精度. 在 BERT 系列模型的启发下,百度提出 ERNIE (enhanced representation through knowledge integration)[7]预训练模型, 进一步增强了预训练模型在中文自然语言处理任务中的表现. ERNIE 通过引入知识图谱和更丰富的语言知识, 能够更有效地捕获汉语中复杂的语义关系和上下文信息, 从而弥补了 BERT 在中文语义理解上的不足. ERNIE 的多层次知识融入策略, 使其在文本分类任务中表现更加优异. 此外, ERNIE 的不同变体也相继推出, 它们在数据规模和知识丰富度上进行了扩展, 适应了更广泛的任务类型. Wang 等人[8]将 ERNIE 预训练模型与 TextRCNN 模型相结合, 显著提升了中文新闻标题的分类准确率, 但其复杂性也相应增加, 在数据量有限情况下会导致模型过拟合. 杨文阳等人[9]利用 ERNIE 和 Bi-LSTM 模型对社交网络文本进行情感分析, 增强了模型在处理情感复杂的社交媒体上的分类性能.

1.2 古文本分类相关工作

文本分类技术已在多个领域得到广泛应用, 如情感分析、医疗文本分析、新闻分类等. 然而在古文本分类领域的应用却相对较少. 这主要是由于古文本具有语言结构复杂、字词异形多样、语义理解困难等特点, 给文本分类模型的构建带来了很大挑战. 在古文本分类领域, Tian 等人[10]利用上下文嵌入模型对中国历史文本进行时期分类, 在古文本分类方面具有良好表现, 但该模型面临着解释性不足和训练测试数据分布不均衡的问题, 限制了其分类性能. 史沛卓等人[11]采用 TextCNN 模型对中国古诗文进行分类, 有效地实现了唐诗、宋词等不同类别的自动分类. 但 TextCNN 模型主要关注文本的局部特征, 未能充分理解整体语义. 汉长安城骨签释文属于古文本, 并且记载着西汉时期的历史信息, 数据繁多, 类别多样并且具备古文本的普遍特点, 但其内容和结构复杂, 包含特殊的语法、词汇和文化背景, 传统方法难以满足准确分类的需求. 近年来的研究揭示了预训练模型与其他文本分类模型结合的巨大潜力, 但在骨签释文的语言结构和用词习惯等方面缺乏足够理解, 并且在面临语义缺失的环境下表现欠佳. 为解决这一问题, 本文提出了一种 EWRCA 骨签释文分类模型, 将文本分类方法应用于汉长安城骨签释文, 有助于深入挖掘和理解这些古代文献的内容. 通过文本分类方法, 可以将释文按照内容进行自动分类. 这不仅提高了古文整理和分析的效率, 还为后续的骨签碎片匹配及文化研究提供了精准的数据支持, 使得对汉长安城历史文化的探索更加系统和深入, 便于实现文物归置及保护. 本研究的主要贡献包括:

(1) 针对关键信息缺失、类别数量较少的骨签释文进行数据预处理与数据增强, 平衡数据类别分布, 增强训练模型的泛化能力;

(2) 针对模型在处理骨签释文语义特征过程中因语义信息丢失所导致的文本表示不充分的问题, 通过选取 ERNIE 模型的 8 层编码器堆叠输出, 融合不同层次的语义特征, 弥补语义信息丢失的影响, 从而增强文本表示的丰富性;

(3) 针对 ERNIE 模型在处理骨签释文中特有词汇时的局限, 引入 Word2Vec 模型训练的词向量来弥补, 并将 ERNIE 模型的 8 层编码器堆叠输出与 Word2Vec 模型生成的词向量构建文本向量融合模块, 为骨签释文特定词汇建立稳定的语义表示, 提高分类精度;

(4) 针对 TextRCNN 模型无法聚焦骨签释文中最具代表性的字符和句子, 从而影响分类精度的问题, 引入融合注意力机制. 其中, 利用多头注意力机制捕捉释文的多层含义和复杂依赖; 在词级和句子级别上利用层次注意力机制的精确加权, 提升关键信息的提取能力, 提高分类准确性.

2 基础理论

2.1 ERNIE 模型

ERNIE 是百度开发的一款高级预训练语言模型, 通过融合大规模知识图谱和深度学习技术, 增强了对语义和语法的理解能力. 在短句子级文本分类中, 将句子作为初始输入, 将词嵌入编码为以词为单位的静态词向量. 将句子嵌入和相应的位置嵌入一起作为 ERNIE 层的输入. ERNIE 层的输入向量形成过程如图 2 所示.

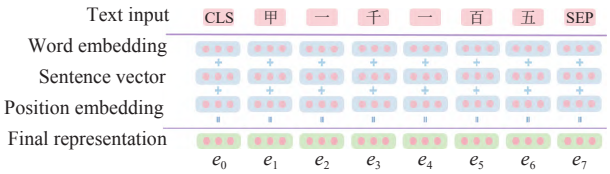

图 2 ERNIE 模型词嵌入过程

图 2 中, [CLS]代表一个句子开头的占位符, 它包含整个句子的信息; [SEP]表示用于区分不同句子的分隔符. 静态词向量$\{e_0, e_1, e_2, \cdots, e_7\}$是通过将词嵌入表示中的输入句子、整句表示和位置表示向量相加作为“甲一千一百五”的输入向量, 然后传递给 ERNIE 层表征生成的. ERNIE 层从输入向量中提取底层的词法和语义信息, 最后生成一个集成上下文的动态词向量表示.

ERNIE 类似于 BERT, 都是基于 Transformer 架构, 利用深度学习技术来理解语言的复杂语义关系, 并采用无监督学习方法在大规模文本数据上进行预训练. 主要区别在于 BERT 的掩码机制采用的 MLM (masked language model), 而 ERNIE 在此基础上引入了知识掩码机制 (knowledge masking). 除了像 BERT 那样进行随机词汇的掩码, ERNIE 还利用大规模知识图谱, 将知识单元 (如实体、关系等) 作为掩码对象, 不仅使模型学习到词汇的语境关系, 还能够捕捉到更深层次的知识结构和语义关联. 骨签释文具有较多的历史背景和专业术语. 因此, ERNIE 比 BERT 更适用于骨签释文分类. BERT 和 ERNIE 的不同掩蔽策略比较如图 3 所示.

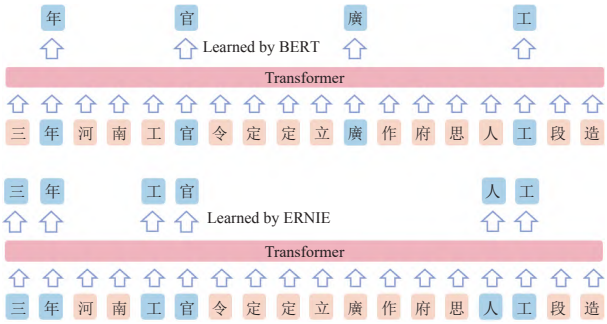

图 3 BERT 和 ERNIE 掩码策略的比较

ERNIE 模型采用多头自注意力机制, 通过并行计算多个注意力头, 在不同子空间中捕捉输入序列的多种特征. 注意自我注意的计算公式如下:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

其中, Q 是查询向量, K 是键向量, V 是值向量, d_k 是键向量的维度. 注意力得分通过点积计算并除以进行缩放, 然后通过 *Softmax* 函数归一化得分, 最后加权求和值向量. 多个头的输出拼接在一起, 经过线性变换, 生成最终的多头注意力输出. 这一机制使模型能够同时考虑序列中的多种特征和长距离依赖关系, 增强对复杂语义的理解能力. 前馈神经网络子层包括两个线性变换和 GELU 激活函数, 对每个词向量进行非线性变换, 进一步丰富表示能力. 每个子层后包含残差连接和层归一化, 缓解梯度消失问题并加速收敛.

2.2 多头注意力机制

多头注意力机制 (multi-head attention mechanism) 是深度学习中广泛应用的注意力机制变体, 最早由 Vaswani 等人[12]在 Transformer 模型中提出. 其主要作用在于增强模型捕捉不同位置间依赖关系的能力, 能够有效提升模型对文本的全局建模能力. 通过多个注意力头从不同角度捕捉词与词之间的远程依赖, 丰富词的全局语义表示. Bi-LSTM 生成的上下文语义在经过多头注意力后, 可以动态调整权重, 突出文本中的关键信息, 尤其在处理复杂的长文本时表现尤为出色.

2.3 层次注意力机制

层次注意力机制 (hierarchical attention mechanism)[13]是一种用于处理文本数据的深度学习方法, 通过对文本进行分层处理, 从而捕捉文本中的重要信息, 具有更强的上下文理解能力. 首先在词级别应用注意力机制, 计算每个词对句子语义的贡献, 通过加权求和得到句子的表示; 接着在句子级别进行类似的操作, 计算各句子对整篇文本的贡献, 从而生成文本的全局语义表示. 该机制不仅能够提高模型对长文本的处理能力, 还增强了对不同层次信息的捕捉与理解能力.

2.4 Bi-LSTM

长短期记忆网络 (LSTM) 能够有效处理长序列数据, 克服传统 RNN 在长距离依赖问题上的不足, 既捕获了输入骨签释文文本特征中的长序依赖关系, 又更好地掌握了输入特征的全局关系. 由于部分骨签释文因断裂而丢失关键信息, 而双向长短时记忆网络 (Bi-LSTM)[14]由两个方向相反的 LSTM 组成, 既能获取正向语义特征信息, 又能获取反向语义特征信息, 通过整合来自文本两端的信息, 模型能够在面对空白或不明确信息时, 完整地重建断裂文本的语义内容. Bi-LSTM 由输入层、前向 LSTM 层、后向 LSTM 层、连接层、输出层组成. 数据被输入到输入层, 按照时间顺序和逆序分别传递到前向和后向 LSTM 层中, 计算每个时间步的隐藏状态, 这意味着输入数据由两个相反方向移动的 LSTM 网络同时处理, 输出层产生的序列受到两个 LSTM 的影响. LSTM 的网络结构如图 4 所示.

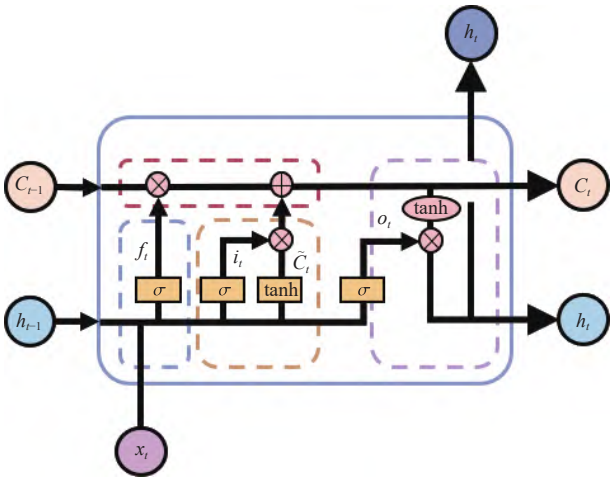

图 4 LSTM 网络结构图

对于给定时间步 t 和输入 x_t, LSTM 的计算过程为:

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{2}$$

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{3}$$

$$\tilde{C}_t=\tanh(W_C\cdot[h_{t-1},x_t]+b_C) \tag{4}$$

$$C_t=f_t\cdot C_{t-1}+i_t\cdot\tilde{C}_t \tag{5}$$

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{6}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{7}$$

其中, f_t、i_t、o_t 分别为遗忘门、输入门、输出门, $\tilde{C}_t$ 表示候选记忆单元的状态, C_{t-1}、C_t 分别表示时间步 t–1 和时间步 t 的记忆单元状态, h_{t-1}、h_t 分别表示时间步 t–1 和时间步 t 的隐藏状态, W 表示权重矩阵, b 为偏置, σ 为 Sigmoid 激活函数.

Bi-LSTM 通过前向和后向两个独立的 LSTM 层同时处理输入序列, 在每个时间步将前向和后向的隐藏状态进行拼接或合并, 全面捕捉序列中的前后文信息, 从而提高模型在应对输入序列的变化和噪声时的稳定性和鲁棒性. Bi-LSTM 网络结构的表达式为:

$$\overrightarrow{h}_t = LSTM(x_t, \overrightarrow{h}_{t-1}) \tag{8}$$

$$\overleftarrow{h}_t = LSTM(x_t, \overleftarrow{h}_{t-1}) \tag{9}$$

$$H_t = [\overrightarrow{h}_t \oplus \overleftarrow{h}_t] \tag{10}$$

在每个时间步 t, $\overrightarrow{h}_t$和$\overleftarrow{h}_t$分别表示前向 LSTM 和后向 LSTM 在时间步 t 的隐藏状态向量; H_t 为 Bi-LSTM 在时间步 t 的输出; $\oplus$表示向量拼接操作.

3 模型构建

本文针对骨签断裂现象导致的骨签释文语义缺失问题提出一种面向语义缺失的骨签释文分类算法——EWRCA 模型, 其由文本向量融合模块与 TextRCNN-MHAtt 文本分类模块相结合. 文本向量融合模块由 ERNIE 多层编码拼接输出的文本向量与 Word2Vec 生成的词向量结合构成, 在处理断裂的骨签释文时, 提供更丰富和更全面的文本表示. TextRCNN-MHAtt 文本分类模块利用 Bi-LSTM 处理整个骨签释文文本序列的信息, 更好地捕捉文本中的时间序列依赖, 并且引入融合注意力机制后可以识别并强调关键文本信息, 增加模型对关键信息的关注度, 并通过池化层保留最显著的特征, 增强模型对文本数据中局部特征的捕捉能力. 并且设置 *Dropout* 正则化机制[15], 防止神经网络的过拟合风险, 经全连接层和分类层将特征映射到最终的输出类别上, 最终得出骨签释文分类结果. EWRCA 模型结构如图 5 所示.

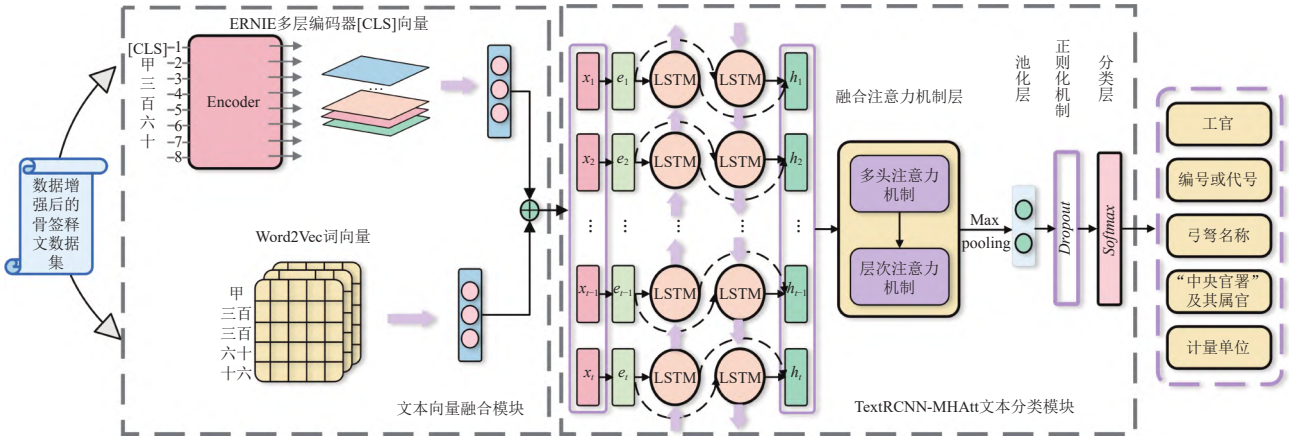

图 5 EWRCA 模型结构图

3.1 数据集构建

实验使用的汉长安城骨签释文数据集是利用基于 ResNet34 的骨签释文识别方法[16], 结合多尺度特征融合策略和焦点损失函数优化, 从骨签图像中提取释文数据后构成. 随后, 通过筛选、去除重复及无用信息, 最终获得 20281 条数据, 分为 5 个类别: 工官类、编号或代号类、弓弩名称类、计量单位类和“中央官署”及其属官类. 这些类别的分布极度不平衡. 汉长安城骨签释文数据集 (部分) 见表 1.

表 1 汉长安城骨签释文数据集 (部分)

汉长安城骨签释文	类别名称
始元四年南陽工官護工卒史令捐丞訢令史意嗇夫猜冗工昌工丁曾造甲	工官
甲二百八十八	编号或代号
乘輿力八石	弓弩名称
六年太僕工繕	“中央官署”及其属官
射三百八十九步	计量单位
第三萬二千七百五十五	编号或代号
服弩力六石	弓弩名称
始元三年河南工官守令石德護工卒史造	工官
力五石二鈞十斤	计量单位

3.1.1 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, 给后续释文分类带来了较大干扰, 故在进行骨签释文分类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: 对骨签释文进行清洗过滤, 删除“/”“…”和“□”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: 对骨签释文进行去停用词处理, 删除诸如“驌”等意义不大的词汇, 减少文本的冗余度.

3.1.2 数据增强

本文所使用的汉长安城骨签释文数据集类别分布存在不均衡性, 类别数量分布如图 6 所示. 如果直接在数据集上应用文本分类算法, 分类精确度会难以提升, 所以需要对骨签释文数据进行增强. 然而, 由于骨签释文具有古文字的独特性, 使用 EDA[17]、回译等数据增强方法容易破坏其语法和语义特性, 造成语义损失或变异. 故采用基于骨签释文规则及模拟骨签断裂方式对弓弩名称类、计量单位类和“中央官署”及其属官类进行数据增强. 骨签释文规则表如表 2 所示. 在模拟骨签断裂过程中, 参考大量真实的断裂骨签, 通过模拟不同断裂位置的骨签表面释文存留情况对释文数据集进行数据增强. 数据增强后的具体数量见表 3. 由于弓弩名称类经过数据增强后的数量依旧少于工官类, 因此采用迭代欠采样的方式将各类数据数量平衡化. 随后将平衡后的骨签数据进行人工标注, 并按 6:2:2 的比例划分为训练集、验证集和测试集.

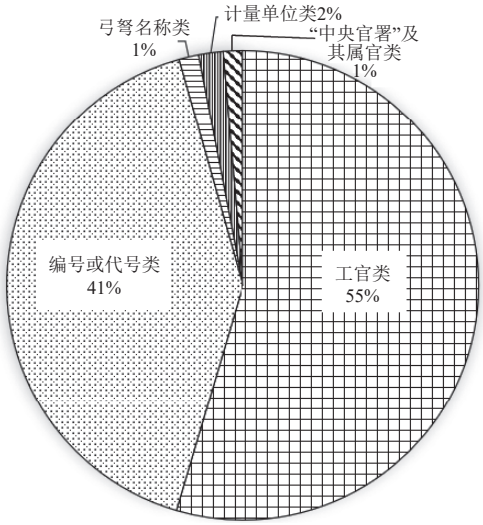

图 6 汉长安城骨签释文类别分布

表 2 弓弩名称类、计量单位类骨签释文规则

汉长安城骨签释文规则	类别名称
“力”+数字 (区间一般在几到十几)+“石”	计量单位
“力”+数字 (一般在十以下)+“石”+数字(数字区间在三十以下)+“斤”	
“力”+数字 (一般在十以下)+“石”+数字 (区间四以下)+“钧”+数字 (区间在三十之内)+“斤”	
“射”+数字 (区间一般在数百)+“步”	
“服”+数字 (一般在十以内)+“石”	弓弩名称
“服”+“力”+数字 (一般在十以内)+“石”	
“服弩”+“力”+数字 (一般在十以内)+“石”	
“大黄”+数字 (一般数值为几十)+“石”	
“大黄”+“力”+数字 (一般数值为几十)+“石”	
“乘舆”+数字 (一般不超过三十)+“石”	
“乘舆”+数字 (一般不超过二十)+“石”+“燥”	
“乘舆”+“燥”+数字 (一般不超过二十)+“石”	
“乘舆”+“御戈”+数字 (一般不超过二十)+“石”	

表 3 数据增强后的类别及其数量

类别名称	数量
工官	11 338
编号或代号	8 534
弓弩名称	7 057
“中央官署”及其属官	10 458
计量单位	12 261

3.2 文本向量融合模块

3.2.1 ERNIE 多层编码

骨签释文具有独特的语言特征, 包括一些与现代汉语意义相同的词汇和大量现代语料中较少出现的专有词汇. ERNIE 模型通过在大规模的现代语料上进行预训练, 学习到语言的通用表示, 在与现代汉语意义相同的骨签释文方面具有出色的理解能力. ERNIE 模型的多层编码器由 12 层相同的结构堆叠而成, 在训练中展现出了强大的语义捕捉能力, 然而常规的编码方式通常只利用最后一层的输出作为文本向量表示, 而骨签释文由于骨签断裂现象造成语义缺失的问题, 仅使用最后一层的输出作为文本向量表示, 会导致在堆叠过程中一些与分类相关的文本语义信息被忽略, 从而影响模型的分类性能. 故本文使用多层编码拼接输出作为文本的向量表示. 多层编码拼接输出是指在多层编码器不断堆叠的同时, 把每一层编码器堆叠得到的[CLS]向量拼接在一起作为最后的输出, 表示最终的文本特征向量. 这种多层编码器拼接输出的方法在面对断裂文本时, 即使某一层的部分信息丢失, 其他层次的信息仍可补充关键语义, 提高模型分类准确性. 但堆叠

过程中过多的层数拼接会产生冗余特征, 导致训练数据的过拟合, 影响分类性能. 因此, 本文选用 8 层编码器的输出[CLS]向量拼接输出, 模型编码器结构如图 7 所示.

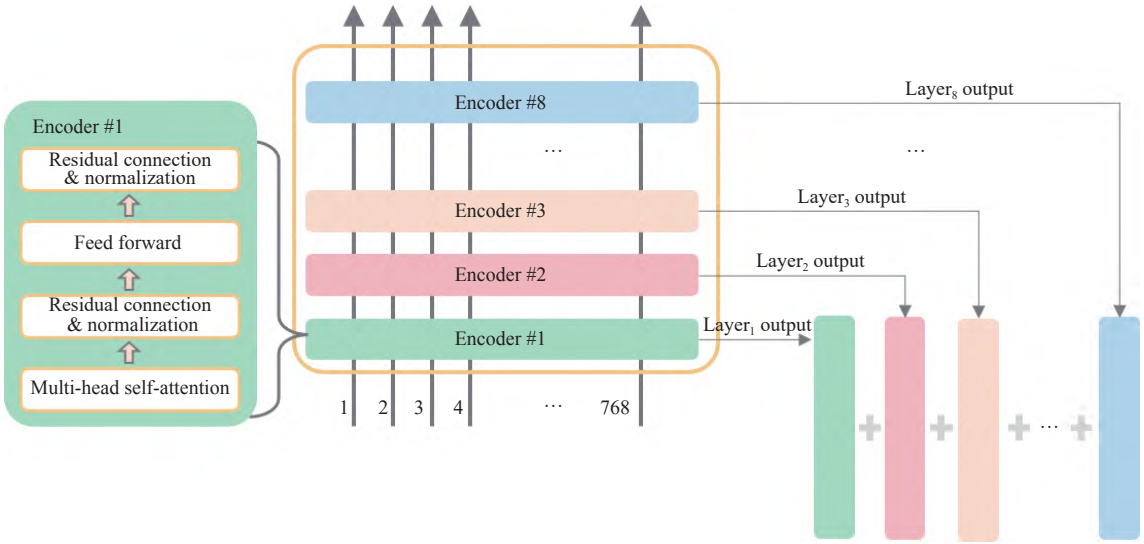

图 7 ERNIE 模型编码层

假设输入序列长度为 n, 每个词向量的维度为 d, 则每一层的输出为 $n\times d$ 的词向量矩阵:

$$Layer_1 \cdot Output = [h_{1,1}, h_{1,2}, \cdots, h_{1,n}] \tag{11}$$

$$Layer_2 \cdot Output = [h_{2,1}, h_{2,2}, \cdots, h_{2,n}] \tag{12}$$

$$\cdots$$

$$Layer_8 \cdot Output = [h_{8,1}, h_{8,2}, \cdots, h_{8,n}] \tag{13}$$

为捕捉不同层次的表示信息, 将第 1–8 层的输出拼接在一起, 使低层关注局部和表层特征, 高层捕捉全局和抽象特征, 拼接后的词向量矩阵表示为:

$$H = [h_{i,1}; h_{i,2}; \cdots; h_{i,8}] \tag{14}$$

其中, i 是序列中词的索引. 这种拼接方式生成的词向量表示维度为原来的 8 倍, 综合了多层编码器的特征信息, 形成更丰富的特征表示.

3.2.2 Word2Vec 词向量

ERNIE 模型在处理骨签释文中的特有词汇和表达时表现出一定的局限性, 为弥补这一不足, 引入了 Word2Vec 模型训练的词向量, Word2Vec[18]是一种基于局部上下文预测的词向量生成方法, 其功能是将文本中的词语映射到一个连续的向量空间. 这种向量表示法能够揭示词语之间的语义相似性及其语法关系. Word2Vec 主要包含两种模型: CBOW (continuous bag of words) 和 Skip-gram. 在 CBOW 模型中, 模型预测目标词语基于其上下文词语, 其结构如图 8 所示. 而在 Skip-gram 中, 模型则基于目标词语来预测其上下文词语. 本研究采用 CBOW 方法训练词向量, 向量维数为 768, 训练后得到的词向量表示为:

$$w_i = [w_{i,1}, w_{i,2}, \cdots, w_{i,768}] \tag{15}$$

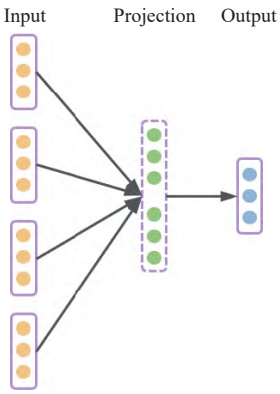

图 8 CBOW 模型结构图

3.2.3 文本向量融合

结合 ERNIE 预训练模型和 Word2Vec 模型的特点, 采用文本向量融合策略, 通过对 ERNIE 多层编码器堆叠生成的[CLS]向量和 Word2Vec 生成的词向量进行维度扩展与降维处理, 并采用特征融合操作, 将两种嵌入整合为统一的文本表示向量, 有效增强了模型

对骨签释文中独有词汇和整体文本语义的表达能力, 从而显著提升了分类和语义分析的性能. 融合后的词向量表示为:

$$v_i = [h_i \oplus w_i] \tag{16}$$

其中, h_i 为使用 ERNIE 模型的第 1–8 层 Encoder 的输出向量, w_i 为使用 Word2Vec 模型的词向量表示, $\oplus$表示向量的拼接融合.

3.3 TextRCNN-MHAtt 文本分类模块

3.3.1 TextRCNN 模型

TextRCNN 模型[19], 结合了循环神经网络 (RNN) 和卷积神经网络 (CNN) 的优势来处理文本数据. 该模型利用 Bi-LSTM 和池化层作为核心组件, 有效捕捉骨签释文中的长距离依赖和局部特征. TextRCNN 模型结构图如图 9 所示.

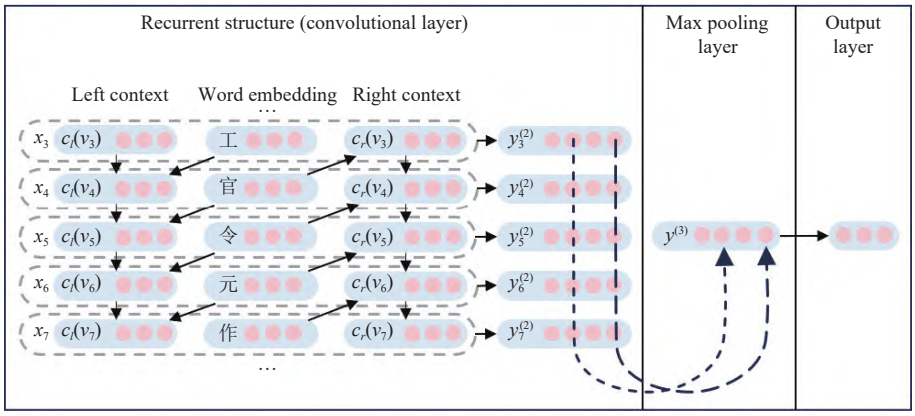

图 9 TextRCNN 模型结构图

图 9 中使用 $c_l(v_i)$ 来定义词 v_i 左边的文本, $c_r(v_i)$ 来定义词 v_i 右边的文本. 其中, $c_l(v_i)$ 和 $c_r(v_i)$ 表示长度为$|c|$的稠密向量.

$$c_l(v_i) = f(W^{(l)}c_l(v_{i+1}) + W^{(sl)}e(v_{i+1})) \tag{17}$$

$$c_r(v_i) = f(W^{(r)}c_r(v_{i+1}) + W^{(sr)}e(v_{i+1})) \tag{18}$$

定义词 v_i 的向量表示:

$$x_i = [c_l(v_i); e(v_i); c_r(v_i)] \tag{19}$$

经过线性变换与 tanh 激活函数后, 得到潜在的语义向量$y_i^{(2)}$, 将每一个语义因素分析, 以确定代表文本的最有用的因素.

$$y_i^{(2)} = \tanh\left(W^{(2)}x_i + b^{(2)}\right) \tag{20}$$

3.3.2 融合注意力机制层

在骨签释文文本中, 短语的不同组成部分可能具有不同意义, 从而对词组含义产生不同影响. TextRCNN 模型难以考虑每次输出的权重, 容易受到全局依赖建模不足的限制, 因此, 本文通过融合多头与层次注意力机制, 增强 TextRCNN 在解析骨签释文数据时的全局依赖建模和局部信息提取能力. 融合注意力机制结构如图 10 所示.

为了提升对全局依赖的建模能力, 本文引入融合注意力机制. 首先通过多个平行的注意力头, 分别从不同的子空间计算词与词之间的相关性, 输出更加丰富的上下文表示, 能够有效应对骨签释文数据中字形复杂、信息分散的情况, 通过全局建模能力捕捉远距离的语义关联, 从而进一步提高模型对复杂文本的理解能力. 引入多头注意力机制不仅弥补了 Bi-LSTM 在长距离依赖建模方面的不足, 还能使模型在文本分类任务中更加精准地提取重要信息, 提升整体性能. 多头注意力机制的计算公式为:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{21}$$

$$Q = W_Q x_i \tag{22}$$

$$K = W_K x_i \tag{23}$$

$$V = W_V x_i \tag{24}$$

其中, W_Q、W_K、W_V 是查询、键和值的线性变换矩阵, d_k 是键向量的维度. 多个注意力头并行执行上述步骤, 形成多头注意力. 通过不同的线性变换, 每个头可以专注于不同维度或位置的信息, 并捕捉到序列中不同部分之间的依赖关系. 然后将所有注意力头的输出拼接在一起, 并通过线性变换得到最终输出:

$$M\text{-}Att(Q,K,V) = Concat(head_1, head_2, \cdots, head_h)W_O \tag{25}$$

其中, $head$ 是注意力头的数量, 每个注意力头独立计算, 最后拼接; W_O 是线性变换矩阵.

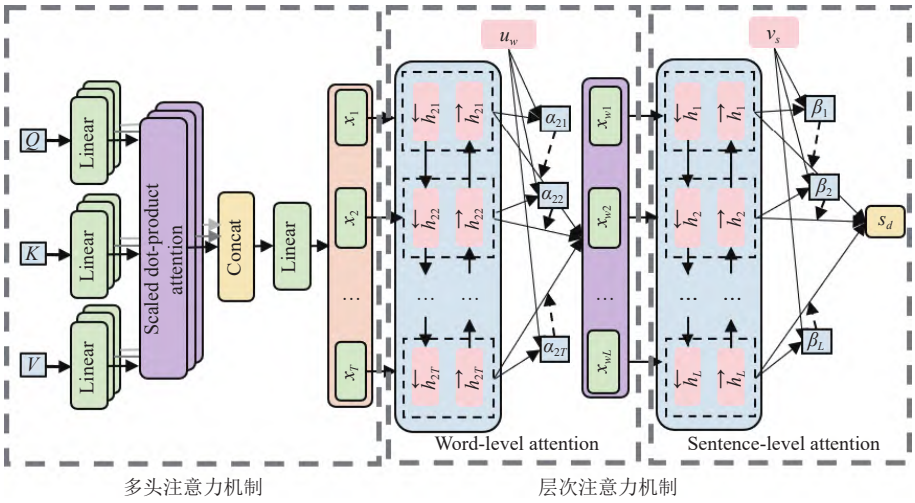

图 10 融合注意力机制结构图

在多头注意力输出的基础上计算词级别的注意力权重, 对字符上下文表示进行加权求和, 得到句子表示:

$$\alpha_i = \frac{\exp\left(u_i^{\mathrm{T}} u_w\right)}{\sum_{j=1}^{n} \exp\left(u_j^{\mathrm{T}} u_w\right)} \tag{26}$$

$$x_w = \sum_{i=1}^{n} \alpha_i \cdot x_i \tag{27}$$

其中, u_w 是可学习的全局上下文向量, α_i 表示每个字符的重要性权重.

在每个句子的表示 x_w 基础上, 应用句子级注意力机制计算句子的权重 β_i, 并对句子表示进行加权求和, 得到整个释文的全局表示 s_d.

$$\beta_i = \frac{\exp\left(v_i^{\mathrm{T}} v_s\right)}{\sum_{j=1}^{m} \exp\left(v_j^{\mathrm{T}} v_s\right)} \tag{28}$$

$$s_d = \sum_{i=1}^{m} \beta_i \cdot x_{w_i} \tag{29}$$

其中, v_s 是可学习的全局上下文向量, β_i 表示每个字符的重要性权重.

对全局文本表示进行非线性变换, 增强模型的表达能力, 使其能够捕捉复杂的非线性关系, 并且在一定程度上减少梯度消失的问题.

$$s_{\mathrm{act}} = ReLU(s_d) \tag{30}$$

3.3.3 池化层

经过池化层对 Bi-LSTM 层输出的文本特征序列进行下采样来降低序列长度和简化数据结构, 从而减少整体模型的计算复杂度并增强模型对骨签释文文本数据中局部特征的捕捉能力. 目前, 最大池化[20]和平均池化是两种最流行的池化技术. 最大池化通过突出最强激活的特征, 强化了模型的判别能力, 这在捕捉关键词或短语时极为有效. 为强化关键特征序列, 通过池化层保留最显著的特征, 增强模型对骨签释文数据中局部特征的捕捉能力, 通过接收的特征序列中提取最重要的特征, 来降低特征的空间维度. 有助于减少序列长度, 简化后续处理流程, 并通过最大值操作确保关键信息的保留. 池化层的表达式为:

$$s_{\mathrm{pool}} = \max(s_{\mathrm{act}}) \tag{31}$$

3.3.4 全连接和分类层

为增强模型在处理未知数据时的泛化性并减少过拟合的可能性, 通过随机丢弃一部分神经元, 防止模型过拟合, 提高模型的泛化能力, 确保骨签释文分类结果的可靠性和精确性.

$$s_{\mathrm{drop}} = Dropout(s_{\mathrm{pool}}, p) \tag{32}$$

其中, p 是保留的比例.

全连接和分类层负责将通过池化层提取并转换的高级特征映射到最终的输出类别上. 对经过 *Dropout* 正则化的特征进行分类或预测. 全连接层将提取的特

征转换为最终的输出概率分布. 全连接层执行如下操作:

$$y_{\text{pred}} = Softmax(W_{\text{fc}} s_{\text{drop}} + b_{\text{fc}}) \tag{33}$$

其中, W_{fc}和b_{fc}分别表示全连接层的权重矩阵和偏差向量. *Softmax* 用于计算每个类别的概率. 可以确保所有输出概率的和为 1, 公式如下:

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{34}$$

使用交叉熵损失函数来衡量预测值和真实标签之间的差异. 通过反向传播算法, 模型根据损失函数的梯度来更新权重和偏差, 以此最小化损失:

$$L = -\sum c_i \log(p_i) \tag{35}$$

其中, L 是损失函数, c_i 为实际标签, p_i 为模型预测的概率.

4 实验结果与分析

4.1 实验环境

实验环境采用 64 位 Windows 10 操作系统, CPU 为 AMD Ryzen 9 5900X 12-Core Processor@3.70 GHz, GPU 为 NVIDIA GeForce RTX 3090, 32 GB 内存. 文本处理软件应用环境为 Python 3.8.16, 并使用 PyTorch[21] 作为深度学习开发框架.

4.2 评价指标

通过准确率、精确率、召回率、*Micro-F*1 以及 *Macro-F*1 分数完成模型性能评估.

● 准确率 (*Acc*, accuracy): 是一个全局评估指标, 表示所有分类正确的预测数量与总预测数量的比例.

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{36}$$

其中, *TP* 为正确预测为正类的样本数; *TN* 为正确预测为负类的样本数; *FN* 为错误地将正类预测为负类的样本数; *FP* 为错误地将负类预测为正类的样本数.

● 精确率 (*P*, precision): 是在所有正类预测中, 实际为正类的比例, 反映模型在预测正类时的准确性.

$$P = \frac{TP}{TP + FP} \tag{37}$$

● 召回率 (*R*, recall): 是在所有实际正类中, 被正确预测为正类的比例, 反映了模型捕获正类样本的能力.

$$R = \frac{TP}{TP + FN} \tag{38}$$

● *Micro-F*1 分数: 是在整个数据集上计算 *F*1 分数, 不区分不同类别.

$$Micro\text{-}F1 = \frac{2 \times P_{\text{micro}} \times R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}} \tag{39}$$

其中, P_{micro}为在所有类别的预测结果合并后, 模型预测为正类的样本中预测正确的比例; R_{micro}表示在所有类别的真实正类样本中, 被模型正确识别出来的比例.

● *Macro-F*1 分数: 首先对每个类别单独计算 *F*1 分数, 然后计算这些 *F*1 分数的平均值. 不考虑每个类别的样本数, 因此每个类别被赋予相同的重要性.

$$Macro\text{-}F1 = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{40}$$

其中, N 为类别数, $F1_i$ 为第 i 个类别的 *F*1 分数.

4.3 实验参数设置

参数的质量将直接影响模型的训练效果, 因此对参数进行调整以优化模型性能. 实验采用 ERNIE-3.0-base-zh 的预训练模型, 损失函数采用交叉熵损失, 模型参数的设置如表 4 所示.

表 4 模型参数设置表

参数	名称	值
ERNIE	预训练模型	ERNIE-3.0-base-zh
D	Word2Vec词向量维度	768
max_len	文本最大长度	38
hidden_size	ERNIE层输出维度	768
hidden_size1	Bi-LSTM隐藏层维度	384
hidden_size2	Bi-LSTM隐藏层维度	192
lr	学习率	1E–5
batch_size	每次迭代使用的样本数	32
epoch	训练批次	20
Dropout	丢弃率	0.5
encoder_layer	ERNIE层数	12

4.4 ERNIE 编码层数对比分析

为验证本文选取 ERNIE 的 8 层编码器输出堆叠的[CLS]向量对骨签数据集分类精度的优越性, 将 ERNIE 不同编码层数的输出用于骨签释文的分类任务进行对比, 实验结果如表 5 所示. 其中, 编码层数为 1 是仅利用最后一层编码器的[CLS]向量作为文本表示, 其他情况则是采用多层编码器的[CLS]向量进行堆叠, 以此作为骨签释文分类任务的输入. 随着编码层数的增加, 模型在准确率、精确率、召回率、*Micro-F*1 和 *Macro-F*1 各项评价指标上均显示出显著提升. 当编码层数增至 8 层时, 各项评估指标均达到最优, 其中准确率提升至 95.20%, 精确率和召回率分别达到 95.62% 和 95.20%. 使用多层编码拼接输出向量表示相对于单层编码能够更全面地捕捉文本上下文语义信息, 从而有效提高模

型分类学习能力和整体分类精度.

如图 11 所示, 在骨签释文数据集上, 模型采用 8 层编码器拼接输出时分类效果最佳, 之后随着层数增加, 性能有所下降. 这是由于过多编码层引入了冗余信息, 从而产生噪声干扰, 影响了分类的准确性. 所以本文模型在经过多次对比实验后选择使用 8 层编码器的拼接输出作为文本向量表示.

表 5 编码层数实验数据表 (%)

编码层数	*Acc*	*P*	*R*	*Micro-F*1	*Macro-F*1
1	89.30	90.90	89.70	89.30	89.43
2	89.82	91.42	90.22	89.82	89.95
3	90.73	92.33	91.13	90.73	90.86
4	91.52	93.12	91.92	91.52	91.65
5	91.86	93.46	92.26	91.86	91.99
6	92.11	93.71	92.51	92.11	92.24
7	93.63	95.23	94.03	93.63	93.76
8	95.20	95.62	95.20	95.20	95.28
9	94.45	94.87	94.45	94.45	94.53
10	93.89	94.31	93.89	93.89	93.97
11	93.33	93.75	93.33	93.41	93.41
12	93.10	93.52	93.10	93.10	93.18

4.5 消融实验与分析

本文在相同骨签释文数据集上完成一组消融实验, 如表 6 所示, 初始模型 TextRCNN 的准确率、精确率和 *Macro-F*1 分数分别为 84.57%、86.56% 和 84.49%. 添加融合注意力机制后, 模型性能得到提升, 表明融合注意力机制能有效捕捉文本中的关键信息. 引入 Word2Vec 词向量后, 增强了模型对骨签释文特有元素的识别和理解能力, 准确率和精确率分别提升了 3.31% 和 1.94%. 进一步整合 ERNIE 预训练模型与 TextRCNN 模型后, 相比单独使用 TextRCNN 模型, 对骨签释文的分类准确率提升了 6.8%. 结合 Word2Vec 词向量和单独注意力机制的 EWRCA 模型在各项指标上都超过了 95%, 并且从图 12 可以看出在单个类别上也有所提升, 充分说明了本文提出方法对骨签释文分类的有效性.

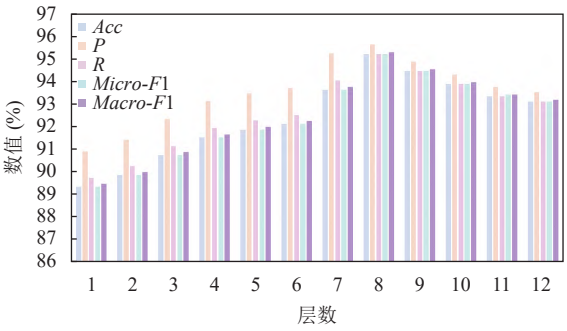

图 11 不同编码层数结果对比图

表 6 消融实验数据表 (%)

模型	*Acc*	*P*	*R*	*Micro-F*1	*Macro-F*1
TextRCNN	84.57	86.56	84.57	84.57	84.49
TextRCNN+MHAtt	85.60	87.59	85.60	85.60	85.50
Word2Vec+TextRCNN	88.91	89.53	89.32	88.91	88.80
Word2Vec+TextRCNN+MHAtt	89.86	91.71	89.86	89.86	89.71
ERNIE+TextRCNN	91.37	91.79	91.72	91.37	91.44
ERNIE+TextRCNN+MHAtt	92.36	92.65	92.66	92.36	92.42
EWRCA	95.20	95.62	95.20	95.20	95.28

4.6 骨签释文分类对比实验与分析

为验证提出的 EWRCA 模型的优越性, 以骨签释文数据集作为实验数据集, 与结合 BERT 几种变体的分类模型 BERT+TextCNN[22]、BERT+DPCNN[23]、BERT+Bi-LSTM[6]、BERT+TextRCNN[24]以及结合 ERNIE 的变体分类模型 ERNIE+TextCNN[25]、ERNIE+Bi-GRU[26]进行对比分析. 实验结果如表 7 所示.

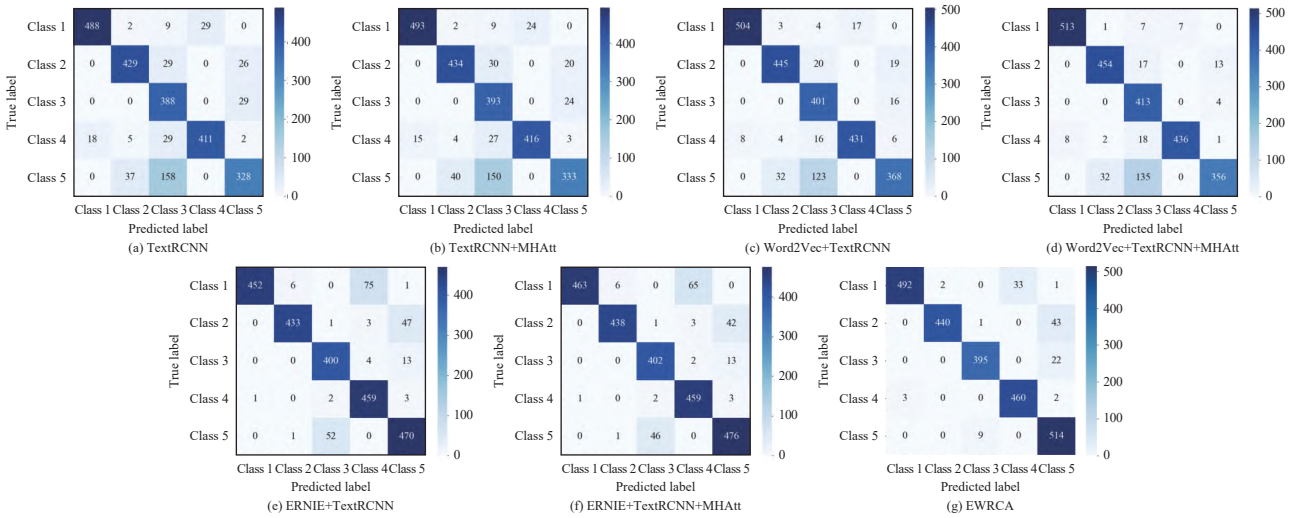

图 12 混淆矩阵图

表 7 对比实验数据表 (%)

模型	*Acc*	*P*	*R*	*Micro-F*1	*Macro-F*1
BERT-Softmax	85.06	86.10	85.06	85.06	85.22
ERNIE-Softmax	86.14	87.02	86.14	86.14	86.26
BERT+TextCNN	87.95	88.30	87.95	87.95	88.01
BERT+DPCNN	88.55	89.25	88.55	88.55	88.66
BERT+Bi-LSTM	54.82	72.79	54.82	54.82	52.27
BERT+TextRCNN	89.02	89.29	89.02	89.02	89.08
ERNIE+TextCNN	89.65	91.08	89.65	89.65	89.59
ERNIE+Bi-GRU	88.26	89.50	88.87	88.26	88.31
EWRCA	95.20	95.62	95.20	95.20	95.28

表 7 数据表明, 本文方法在一定程度上提高了文本分类的精确度, 分类效果均优于其他算法. 在骨签释文特殊文本分类任务中, EWRCA 模型能更好地捕捉和理解文中的深层语义. 由于骨签释文的特性, ERNIE-Softmax 模型在各项性能指标上均优于 BERT-Softmax, 且结合 ERNIE 预训练模型的 ERNIE+TextCNN 模型与结合 BERT 变体的 BERT+TextCNN 模型相比在准确率和精确率上分别提高了 1.7% 和 2.78%, 显示了使用 ERNIE 预训练模型在增强模型对骨签释文特征理解和提取方面的有效性. EWRCA 在 *Macro-F*1 值上相较 ERNIE+TextCNN、ERNIE+Bi-GRU 分别提高了 5.69%、6.97%, 表明其在处理复杂文本时的优越性能.

为进一步直观地展示 EWRCA 模型在骨签释文分类任务的优越性, 分析每个模型的训练过程, 各模型训练过程 Loss 曲线如图 13 所示.

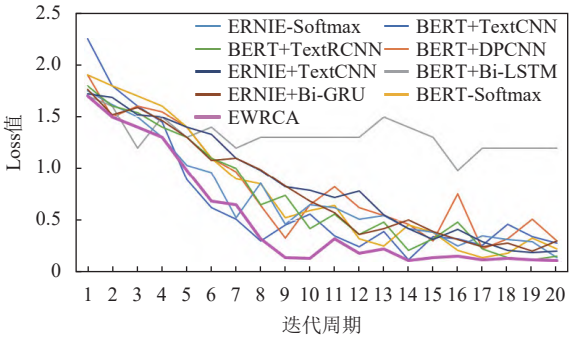

图 13 Loss 曲线对比图

结果显示, EWRCA 模型的损失值在训练的前 5 个周期内迅速下降, 并在第 5 个周期之后趋于平稳, 表明该模型具有较快的收敛速度. 与 BERT+TextCNN 和 ERNIE+Bi-GRU 等模型相比, 尽管这些模型在初期损失值下降较快, 但在随后的训练周期中, 其损失值表现出较大的波动, 显示出这些模型在骨签释文数据集上的长期稳定性不足. 同时, ERNIE+TextCNN 模型在第 10 个周期之后才开始表现出稳定的下降趋势, 其收敛速度显著较慢. 总体而言, EWRCA 模型的表现优于其他模型, 不仅在初期损失降低速度上更为显著, 而且在后续训练中保持了更高的稳定性. 通过实验验证了本文模型在性能上具有更优异的表现.

5 结论

本文提出的 EWRCA 模型是一种针对样本类别数量分布不均、部分文本信息缺失的骨签释文分类模型, 旨在解决断裂骨签存在语义信息缺失导致分类准确率较低的问题. 本文通过基于规则的方法对骨签释文进行数据增强以平衡样本数量; 利用 ERNIE 的 8 层编码器拼接输出的[CLS]向量, 帮助模型捕捉复杂的语义依赖关系和句子级别的特征; 通过 Word2Vec 训练的词向量提供丰富的语言信息和词汇的语义表达, 补充了 ERNIE 的[CLS]向量可能忽略的词级特征和细粒度语义, 增强模型对古文特有元素的识别和理解能力; 随后由 Bi-LSTM 层进一步处理时间序列数据捕捉前后文的依赖性, 引入融合注意力机制关注骨签释文信息中的关键部分, 提高模型对局部特征的区分能力并通过池化层确保分类结果的稳定性和高效性. 实验结果表明, 与几种常见的分类模型相比, 本文方法在骨签释文分类任务中的精确率和准确率达到 95.62%、95.2%, 其他指标均达到 95% 以上, 可以辅助专家学者提高对骨签释文研究的工作效率.

本文选择在图像识别后的文本分类阶段展开研究, 在数据预处理和构建阶段耗费了较多的人力, 未来计划进一步探索从图像识别角度直接开展文本分类的可能性, 并尝试结合图像识别与文本分类的多模态方法, 以降低人力成本, 同时挖掘骨签释文的更多潜在价值.

参考文献

1 高杰. 汉长安城遗址出土骨签名物和用法再议. 华夏考古, 2011(3): 109–113, 149.
2 卢烈炎. 汉长安城未央宫出土骨签初步研究 [硕士学位论文]. 西安: 西北大学, 2013.
3 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: ACL, 2019. 4171–4186.

4 Yu BH, Deng C, Bu LP. Policy text classification algorithm based on BERT. Proceedings of the 11th International Conference of Information and Communication Technology (ICTech). Wuhan: IEEE, 2022. 488–491.

5 Cui YR, Huang CB. A Chinese text classification method based on BERT and convolutional neural network. Proceedings of the 7th International Conference on Systems and Informatics (ICSAI). Chongqing: IEEE, 2021. 1–6.

6 郝婷, 王薇. 融合 BERT 和 BiLSTM 的中文短文本分类研究. 软件工程, 2023, 26(3): 58–62.

7 Sun Y, Wang SH, Li YK, *et al*. ERNIE: Enhanced representation through knowledge integration. arXiv: 1904.09223, 2019.

8 Wang Q, Li X. Chinese news title classification model based on ERNIE-TextRCNN. Proceedings of the 5th International Conference on Machine Learning and Natural Language Processing. Sanya: ACM, 2022. 147–151.

9 杨文阳, 孔科迪. 基于 ERNIE-BiLSTM 的社交网络文本情感分析. 中国电子科学研究院学报, 2023, 18(4): 321–327. [doi: 10.3969/j.issn.1673-5692.2023.04.004]

10 Tian ZY, Kübler S. Period classification in Chinese historical texts. Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Punta Cana: ACL, 2021. 168–177.

11 史沛卓, 陈凯天, 钟叶珂, 等. 基于 TextCNN 的中国古诗文分类方法研究. 电子技术与软件工程, 2021(10): 190–192.

12 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

13 Yang ZC, Yang DY, Dyer C, *et al*. Hierarchical attention networks for document classification. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016. 1480–1489.

14 Zhang S, Zheng DQ, Hu XC, *et al*. Bidirectional long short-term memory networks for relation classification. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. Shanghai: ACL, 2015. 73–78.

15 Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958.

16 Du MX, Wang HQ, Liu R, *et al*. Research on bone stick text recognition method with multi-scale feature fusion. Applied Sciences, 2022, 12(24): 12507. [doi: 10.3390/app122412507]

17 Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 6382–6388.

18 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

19 Lai SW, Xu LH, Liu K, *et al*. Recurrent convolutional neural networks for text classification. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin: AAAI, 2015. 2267–2273.

20 Yu DJ, Wang HL, Chen PQ, *et al*. Mixed pooling for convolutional neural networks. Proceedings of the 9th International Conference on Rough Sets and Knowledge Technology. Shanghai: Springer, 2014. 364–375.

21 Paszke A, Gross S, Massa F, *et al*. PyTorch: An imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: ACM, 2019. 721.

22 Zhang JW, Li L, Yu B. Short text classification of invoices based on BERT-TextCNN. Proceedings of the 2024 International Conference on Artificial Intelligence and Communication Technology. Singapore: Springer, 2024. 153–164.

23 张静, 高子信, 丁伟杰. 基于 BERT-DPCNN 的警情文本分类研究. 数据分析与知识发现, 2024, 1–15. http://kns.cnki.net/kcms/detail/10.1478.G2.20240313.1318.008.html. (2024-03-13)[2024-11-26]

24 Yuan SJ, Wang QX. Imbalanced traffic accident text classification based on BERT-RCNN. Journal of Physics: Conference Series, 2022, 2170(1): 012003. [doi: 10.1088/1742-6596/2170/1/012003]

25 Wang MT, Xu JW. Research on Chinese short text classification based on ERNIE-TextCNN model. Proceedings of the 3rd International Conference on Advanced Algorithms and Neural Networks. Qingdao: SPIE, 2023. 315–319.

26 常俊豪, 武钰智. 基于 ERNIE_BiGRU 模型的中文医疗文本分类. 电脑知识与技术, 2022, 18(1): 101–104. [doi: 10.3969/j.issn.1009-3044.2022.1.dnzsyjs-itrzyksb202201037]

(校对责编: 王欣欣)