

面向破损古籍的全局性语义感知混合修复模型

王惠芳, 钟志峰, 王君怡, 彭宅琨, 宋宇凡, 黄培沛

(湖北大学人工智能学院, 湖北 武汉 430062)

摘要: 作为中华优秀传统文化重要组成部分的中文古籍, 由于保存条件限制、岁月侵蚀甚至人为损毁等原因, 造成部分古籍出现不同程度的损坏和缺失。针对此问题, 提出一种面向破损缺失古籍的语义感知混合修复模型。首先, 通过Transformer处理输入文本序列, 以捕获丰富的特征表示; 其次, 采用Bi-LSTM网络从序列特征表示中提取顺序信息和双向上下文信息, 进一步增强模型对上下文的理解能力; 最后, 引入全局注意力, 计算文本缺失位置的上下文与缺字位置的关联度, 动态调整不同位置的信息权重。通过以上处理的修复模型能较好地实现缺失文本修复。对比实验表明, 该修复模型在公开数据集《水浒传》上的准确率显著优于单一的语言模型, 训练集准确率从90.95%上升至95.97%, 测试集准确率从54.62%上升至71.58%。同时, 该模型直接用于修复医书古籍《神农本草经》, 性能依然较为优异, 说明该模型具备良好的泛化能力。

关键词: 文本修复; 语义感知; 混合模型; Transformer; Bi-LSTM; 全局注意力

DOI: 10.11907/rjdk.251359

扫描二维码阅读全文:



中图分类号: TP18

文献标识码: A

文章编号: 1672-7800(2026)009-0050-08

A Global Semantic Awareness Hybrid Restoration Model for Damaged Ancient Texts

WANG Huifang, ZHONG Zhifeng, WANG Junyi, PENG Zhaikun, SONG Yufan, HUANG Peipei

(School of Artificial Intelligence, Hubei University, Wuhan 430062, China)

Abstract: As an important part of Chinese cultural heritage, some ancient books in Chinese have been damaged and missing to varying degrees due to restrictions on preservation conditions, erosion over time, and even man-made damage. In order to solve this problem, this paper proposes a semantic perception hybrid restoration model for damaged and missing ancient books. The model first processes the input text sequence with Transformer to capture rich feature representations. Then, the Bi-LSTM network is used to extract the sequence information and two-way context information from the sequence feature representation to further enhance the model's ability to understand the context. Finally, global attention is introduced to calculate the correlation between the context of the missing position of the text and the position of the missing word, and the information weight of different positions is dynamically adjusted. Through the above processing of the repair model, the repair of missing text can be better realized. Comparative experiments show that the accuracy of the repair model on the public dataset "Water Margin" is significantly better than that of a single language model, and the accuracy of the training set increases from 90.95% to 95.97%, and the accuracy of the test set increases from 54.62% to 71.58%. At the same time, the model is directly used to restore the ancient medical book "Shen-nong's Materia Medica", and the performance is still excellent, indicating that the model has good generalization ability.

Key Words: text fixing; semantic perception; hybrid model; Transformer; Bi-LSTM; global attention

收稿日期: 2025-07-16

基金项目: 国家自然科学基金项目(62273135)

作者简介: 王惠芳(2001-), 女, CCF 学生会会员, 湖北大学人工智能学院硕士研究生, 研究方向为自然语言处理; 钟志峰(1971-), 男, 博士, 湖北大学人工智能学院教授、博士生导师, 研究方向为自然语言处理、信息处理、系统集成; 王君怡(2000-), 女, 湖北大学人工智能学院硕士研究生, 研究方向为自然语言处理; 彭宅琨(2002-), 男, 湖北大学人工智能学院硕士研究生, 研究方向为图像识别; 宋宇凡(2001-), 女, 湖北大学人工智能学院硕士研究生, 研究方向为图像描述; 黄培沛(2001-), 男, 湖北大学人工智能学院硕士研究生, 研究方向为自然语言处理。本文通讯作者: 钟志峰。

0 引言

中文古籍著作作为中华文化的重要组成部分,蕴藏着丰富的历史、文化和文学价值^[1]。然而,由于保护条件有限、保存技术水平限制、岁月侵蚀,甚至人为损毁等原因,造成许多著作在流传过程中出现了不同程度的文本损毁^[2]。这不仅使得文字内容的释读和理解变得困难,更是对古籍研究提出了极大挑战。而古籍文本修复是文物修复的一部分,古籍文本修复水平提升也可提高文物保护技术和保存质量,对文物保护工作具有积极作用^[3]。因此,如何对古籍文字进行有效修复,成为当前人文领域的一大研究热点。

传统的古籍修复是专业科研人员根据古籍中其他部分的内容及字形,结合历史背景、语境和古文字规律,对缺失或模糊的字词进行合理推测和补充^[5]。该方法容易受到修复者自身知识、经验、认知等因素影响,导致修复过程中出现人为错误,并且人工修复需要大量的时间和精力,尤其是对于内容繁多、字形复杂的古籍,修复工作可能需要较长时间才能完成,效率较低。近年来,深度学习技术的发展为古籍文字修复提供了新的契机。Brown^[6]提出基于词类别的N-gram模型,通过历史词序列预测当前词,讨论了从文本样本以前的单词中预测单词的问题。但是,随着N值的增加,数据稀疏性随N值增大而显著,未出现的字词组合无法处理,且N-gram模型只考虑固定窗口内的上下文,而忽略了句子中词语之间的远距离依赖^[7]。为了克服N-gram模型在捕捉长距离依赖上的局限性,Elman^[8]提出循环神经网络RNN的语言模型,它能够解决N-gram的长距离依赖缺陷。但是,RNN在处理长文本时容易出现梯度消失问题,导致模型无法有效记住较早的词语信息。为了解决梯度消失问题,Hochreiter等^[9]提出长短期记忆网络模型(Long Short-Term Memory, LSTM),LSTM引入门控机制控制信息流,解决了传统RNN的长期依赖问题,但是LSTM只能单向处理序列,缺乏对后续时刻信息的利用。针对此问题,Graves等^[10]提出了Bi-LSTM模型(Bi-directional Long Short-Term Memory),Bi-LSTM可以融合前后文信息,使得模型能够同时考虑过去和未来信息,从而弥补LSTM只能单向处理序列的缺陷。但是,Bi-LSTM序列计算无法并行化,效率受限。近年来,随着Transformer与大语言模型推动修复范式变革,Devlin等^[11]提出BERT模型,该模型可以很好地解决Bi-LSTM无法并行计算的问题。BERT模型和Liu等^[12]提出的基于Bert的预训练模型RoBERTa在民间文学文本修复和中医古籍修复等方面均取得了不错的效果。但是,在使用RoBERTa预训练模型进行中医古籍修复的任务中,其虽然能够较为准确地预测出缺失字词,但测试集准确率很低,而且一句中只能进行一个字的预测^[14]。曹熊能等^[13]在基于民间文学文本修复中提

出的BERT结合知识蒸馏方法,能够对句子中多个位置多个字词进行修复,但是与原文相对照的正确率依旧较低。田文靖^[15]提出TFM-PNSFL方法,使填充缺失部分很好地匹配上下文相对应的语义,但该方法适用于英文文本预测填充,未能对缺失的古籍文本进行字符级和词级修复。段荧等^[16]提出一种基于U型网络框架和部分卷积运算的文字图片修复模型,根据已有笔画细节对古籍污损部分进行修复,并对古籍污损汉字的字体形状和笔画走向的细部特征进行复原。刘栋斌等^[17]提出一种基于内容风格迁移的残缺稀疏文字图像盲修复方法,利用循环生成对抗网络构建修复前后文字图像间的全局关联像素信息,将残缺古籍文字的图像内容风格特征迁移为完整文字图像从而进行修复,但是必须有图像参与,不能够依靠字词上下文进行修复。郑淑婷^[18]提出一种基于关键语义信息选择机制的文本填充模型,在Yelp英文数据集上的填充示例语义连贯且符合原文语义,但在古籍文本数据集上填充效果不佳。Vaswani等^[19]提出Transformer模型,其使用自注意力机制,能够在一个句子中的所有词之间建立依赖关系,轻松处理较长距离依赖,并且能够实现并行计算,但在处理短距离依赖或局部特征时不如Bi-LSTM高效。

综上所述,本文提出了一种面向破损中文古籍文本的语义感知混合修复模型,旨在有效提升缺失字词预测与还原能力。该模型以Transformer编码器为基础框架,融合了Bi-LSTM与全局注意力机制(Global Attention),充分发挥了多层次语义建模能力。其中,Transformer编码器擅长建模长距离依赖关系,Bi-LSTM通过门控结构增强对局部上下文的感知能力,而全局注意力机制则有助于精确聚合关键语义信息,从而提高了整体修复质量。基于此设计,构建了融合全局与局部特征的语义感知修复网络,用于对中文古籍中的缺失字符进行上下文相关性预测与填补。在实验设置中,选取中国古典小说《水浒传》作为主要训练与评估语料,模型在训练集上达到了95.97%的预测准确率,在测试集上准确率也达到71.58%。此外,为验证模型泛化能力,本文将其直接应用于医药类古籍《神农本草经》的修复任务中,模型依然展现出较优性能。上述结果表明,所提混合模型在传统古籍修复场景中具有较强的语义理解能力和泛化性。

1 相关算法

1.1 Transformer 编码器

Transformer编码器包含多头自注意力(Multi head Attention)机制层和全连接前馈层,结构如图1所示^[20]。自注意力机制是Transformer中多头自注意力的组成部分,如图2所示。对应的输入向量矩阵 X 通过线性层分别转换为查询矩阵(Q)、键矩阵(K)和值矩阵(V)计算点积,得到相似性得分,然后通过Softmax函数归一化,得到注意力权重,

如式(1)、式(2)所示。

$$Q = XW_Q, K = XW_K, V = XW_V \quad (1)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

其中, W_Q, W_K, W_V 是学习到的权重矩阵, $\sqrt{d_k}$ 是向量维度。

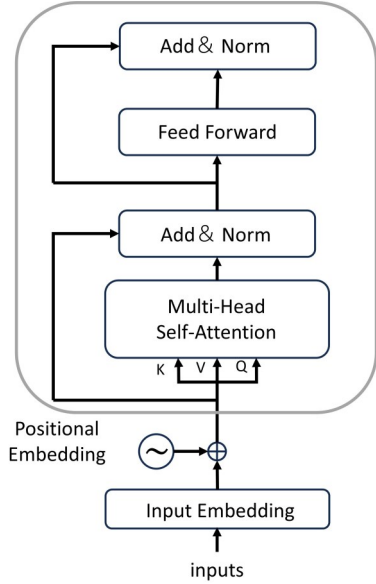


Fig. 1 Transformer encoder structure

图1 Transformer 编码器结构

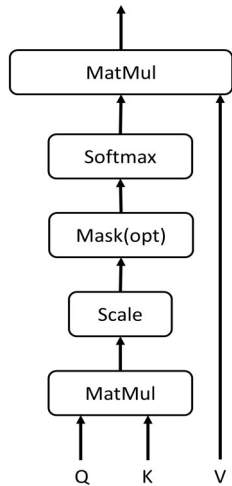


Fig. 2 The self-attention mechanism structure

图2 自注意力机制结构

(1)多头自注意力机制子层。Transformer所使用的多头注意力机制是由多个自注意力机制组合形成,如式(3)、式(4)所示。

$$head_i = Attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right)V_i \quad (3)$$

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_h)W_o \quad (4)$$

其中,下标 h 表示多头注意力机制中头的个数, Q, K, V

和 $\sqrt{d_k}$ 表示与自注意力机制中一致。

(2)全连接前馈网络子层。该子层由两个线性变换构成,其中加入了 ReLU 激活函数,如式(5)所示。

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (5)$$

其中, W_1, W_2 和 b_1, b_2 分别是线性变换的权重和偏置。

1.2 Bi-LSTM 网络结构

LSTM 的 3 个门控机制分别是遗忘门、输入门和输出门,它们通过控制信息流动,从而保留长期序列信息,如图 3 所示。

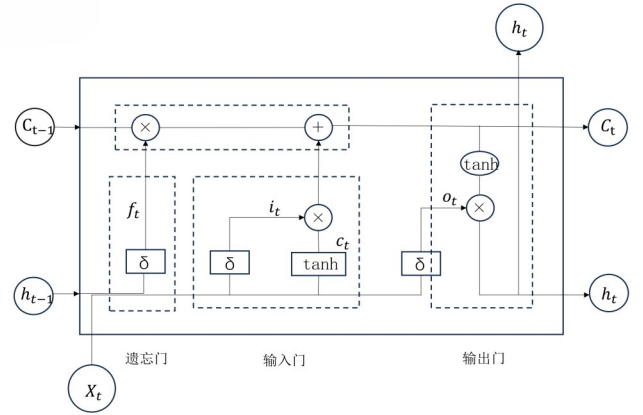


Fig. 3 Structure of LSTM cells

图3 LSTM单元结构

遗忘门决定哪些信息需要被保留,接收当前输入 x_t 和上一时刻 h_{t-1} 的隐藏状态作为输入,通过一个 Sigmoid 激活函数将其映射到 0~1 区间的值,如式(6)所示。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (6)$$

其中, f_t 表示遗忘门, x_t 表示当前输入, W_f 表示遗忘门的权重矩阵, b_f 表示偏置向量, h_{t-1} 表示前一时间步的隐状态。 σ 表示 Sigmoid 激活函数; $[\cdot, \cdot]$ 表示拼接操作。

输入门控制当前输入中有多少信息被更新,通过 Sigmoid 函数计算出更新比例,同时通过 Tanh 激活函数对当前输入进行变换,得到需要更新到细胞状态中的信息,如式(7)和式(8)所示。

$$i_t = (W_i \cdot [h_{t-1}, x_t] + b_i) \quad (7)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (8)$$

其中, i_t 表示输入门的输出, W_i 表示输入门的权重矩阵, b_i 表示输入门偏置项, $\tilde{C}_t$ 表示候选的记忆单元状态, W_c 表示记忆状态的权重矩阵, b_c 表示记忆状态的偏置项。

输出门决定哪些信息将被输出作为当前时刻的隐藏状态,它接收上一时刻的隐藏状态和当前时刻的输入,并将其作为输入,通过 Sigmoid 函数计算输出比例,再将其与经过 Tanh 激活函数处理后的细胞状态相乘,得到当前时刻的隐藏状态,如式(9)一式(11)所示。

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (9)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (10)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (11)$$

其中, o_t 表示输出门的输出; W_o 表示输出门的权重矩阵; b_o 表示输出门偏置项, C_t 表示当前时刻的记忆单元状态, h_t 表示当前时刻的隐藏状态。

Bi-LSTM 包含两个独立的 LSTM, 一个是从前往后处理序列的正向 LSTM, 另一个是从后往前处理序列的反向 LSTM, 考虑了序列中每个时间步前后的上下文信息^[21]。Bi-LSTM 结构示意图如图 4 所示。

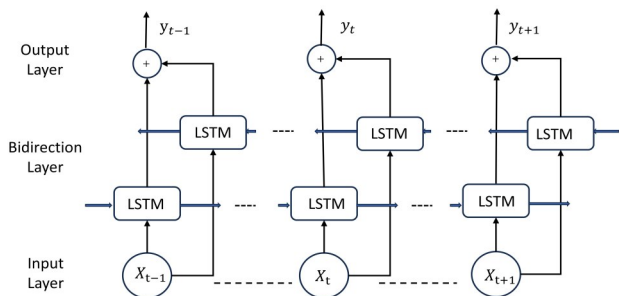


Fig. 4 Schematic diagram of Bi-LSTM structure

图 4 Bi-LSTM 结构示意图

1.3 全局注意力机制

Bahdanau 等^[22]提出全局注意力机制, 该机制可以聚焦于全序列中所有位置的相对关系, 从而加强对长距离依赖的建模。全局注意力机制是指在处理某一时刻的信息时, 模型能够考虑到所有输入数据(全局范围)以进行决策。全局注意力通常会根据输入的所有特征计算一个加权值, 通过计算一个权重矩阵给不同的输入部分分配重要性权重, 从而决定哪些部分的信息应该被重点关注, 哪些部分可以忽略, 如图 5 所示。

其中, 第 n 个字词在句子中的权重 b_n , 是根据 n 时刻的

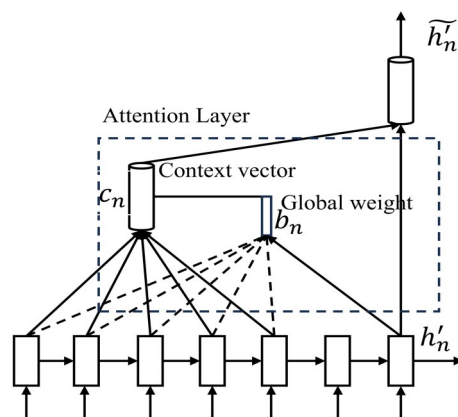


Fig. 5 Schematic diagram of global attention structure

图 5 全局注意力结构示意图

h_n 和 h'_t 计算而来, 权重计算公式如式(12)所示。

$$b_n = \text{align}(h'_n, h_n) \quad (12)$$

其中, h'_n 代表最后的 hidden state, h_n 表示 n 时刻的 hidden state。

2 中文古籍文本修复模型

为解决古籍著作在流传过程中文本字词出现不同程度的损坏和缺失问题, 本文提出了一种面向破损古籍的全局性语义感知混合修复模型。该模型着重对文本上下文的语境理解和多层次的文本特征融合方法进行了创新, 旨在高效地处理序列建模任务, 有效地预测文本所缺失字词。该模型通过 Transformer 编码器、Bi-LSTM 层进行多层次特征提取, 之后采用全局注意力机制计算文本各部分与缺失字位置的关联度, 动态调整不同位置的信息权重, 如图 6 所示。

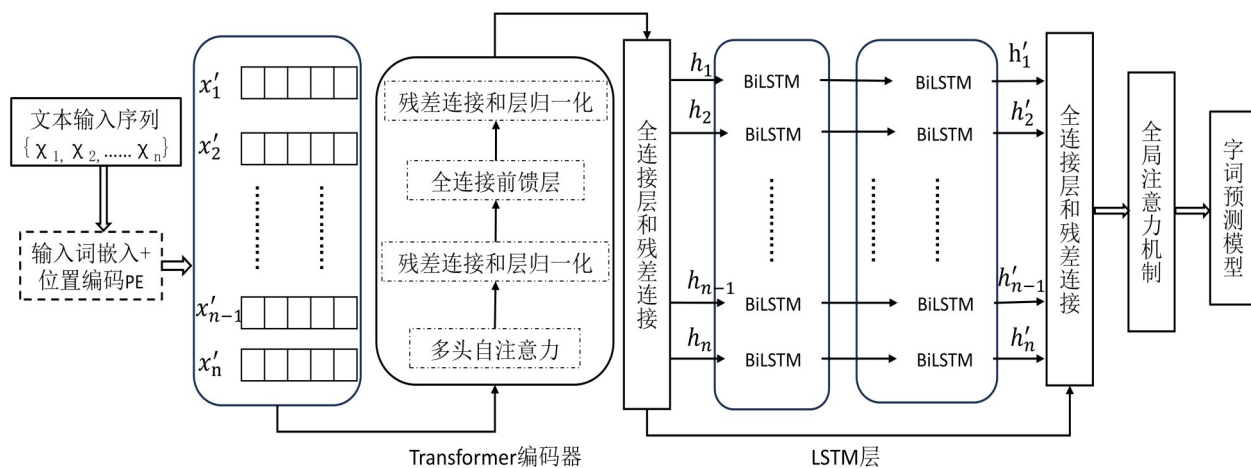


Fig. 6 The hybrid repair model structure

图 6 混合修复模型结构

2.1 字词嵌入层

Transformer 结构中引入了位置编码, 用于捕捉序列中不同位置的关系。位置编码是为了解决其核心组件自注意力机制的位置不敏感性而设计的创新方案, 补偿了

Transformer 由于完全依赖注意力机制而可能失去的顺序信息捕捉能力。位置编码通过正弦和余弦函数构成, 这样 Transformer 能够有效处理不同长度的序列, 捕捉不同频率的信息, 如式(13)、式(14)所示。

$$PE_{pos, 2i} = \sin\left(\frac{pos}{10\,000^{\frac{2i}{d_{model}}}}\right) \quad (13)$$

$$PE_{pos, 2i+1} = \cos\left(\frac{pos}{10\,000^{\frac{2i}{d_{model}}}}\right) \quad (14)$$

其中, PE 表示位置编码, pos 表示字在句子中的位置, d 表示位置编码的维度, i 代表维度指数, $2i$ 表示偶数维度, $2i+1$ 表示奇数维度。

文本数据输入序列 $X = \{x_1, x_2, \dots, x_{n-1}, x_n\}$, 通过字符嵌入 $E_{token} = \{e_1, e_2, \dots, e_{n-1}, e_n\}$ 和位置编码 $E_{pos} = \{e'_1, e'_2, \dots, e'_{n-1}, e'_n\}$ 相加得到 Transformer 编码器的真正输入序列 $X' = \{x'_1, x'_2, \dots, x'_{n-1}, x'_n\}$, 如式(15)所示。

$$X' = E_{token} + E_{pos} \quad (15)$$

其中, E_{token} 表示字符嵌入, E_{pos} 表示位置编码。

2.2 特征提取层

本文选用的是 Transformer 编码器, 对于输入的序列 $X' = \{x'_1, x'_2, \dots, x'_{n-1}, x'_n\}$, 每一层都在提取更高级别的特征。通过逐层处理, 模型能够组合低层特征以形成更复杂的表示, 从而提高语义理解能力, 这对于字词预测任务尤其重要, 它需要模型理解上下文中的多层次含义。Transformer 编码器通过自注意力机制, 能够捕捉输入序列中任意两个字之间的关系, 无论是对长距离还是短距离的依赖, 这使得模型能够更好地理解字词在上下文中的作用。将通过 Transformer 编码器处理后的这些序列经过全连接层, 再经过残差连接得到序列 $H = \{h_1, h_2, \dots, h_{n-1}, h_n\}$, 如式(16)所示。该残差网络是 Transformer 输出和映射后的初始嵌入构成, 以保留初始嵌入信息, 防止信息在 Transformer 编码器的深层变换中丢失。

$$Transformer'_{out} = Transformer_{out} + self.f_c(X') \quad (16)$$

其中, $self.f_c(X')$ 表示字符初始嵌入映射, $Transformer_{out}$ 表示编码器的输出, $Transformer'_{out}$ 表示残差后的输出。

2.3 语义提取层

本文语义提取层采用的是 Bi-LSTM, 在 Transformer 编码层后连接两层 Bi-LSTM 网络, 将残差连接后的输出序列 $H = \{h_1, h_2, \dots, h_{n-1}, h_n\}$ 输入到 Bi-LSTM 模块。Bi-LSTM 网络将 Transformer 编码器处理后的序列 $H = \{h_1, h_2, \dots, h_{n-1}, h_n\}$ 进一步进行特征提取, 特别是 Transformer 编码器中未提取到的局部特征, 同时也能够更好地防止模型过拟合。Bi-LSTM 的引入在局部和短期依赖关系上提供了补充, 使得模型在处理复杂序列时更加稳健。其引入更细腻的局部上下文信息, 加强了序列顺序和前后依赖的捕捉能力, 得到了更加充足的上下文信息输出序列 $H' = \{h'_1, h'_2, \dots, h'_{n-1}, h'_n\}$ 。将经过 Bi-LSTM 层后的输出序列经过全连接层, 再经过一个残差连接, 得到 H'' , 该残差网络是由 LSTM 输出和 Transformer 残差输出构成, 如式

(17)、式(18)所示。

$$H'' \in R^{T \times B \times d} \quad (17)$$

$$BiLSTM'_{out} = BiLSTM_{out} + Transformer'_{out} \quad (18)$$

其中, B 为批大小, T 为序列长度, $d=2 \cdot \text{lstm_hidden_dim}$, $Transformer'_{out}$ 表示第 1 个残差的输出, $BiLSTM_{out}$ 表示经过 Bi-LSTM 的输出, $BiLSTM'_{out}$ 表示第 2 个残差的输出。

2.4 分类预测层

将残差后的数据 H'' 经过一个全局注意力机制层, 与标准的 Transformer 编码器堆叠通过多头自注意力层处理顺序元素不同, 这一全局注意力层作用于整个编码序列输出, 通过强制对字符依赖关系进行全局视角的处理, 补充了编码器的局部层次处理。全局注意力层的引入, 通过有效整合 LSTM 输出并聚焦于更重要的信息, 这种设计提升了模型针对中文字词预测任务的表现。它不仅提高了模型准确性, 还增强了语句上下文理解能力。最后一个时间步 O_{final} 的输出, 需经过全局注意力层, 将 O_{final} 输入全连接层进行分类, 从而得到最终字词预测模型, 如式(19)所示。

$$Output = \text{Softmax}(O_{final}W_c + b_c) \quad (19)$$

其中, W_c 为权重矩阵, b_c 为偏置。

3 实验设置

3.1 数据集处理与实验环境

本文对比实验使用的数据集是《水浒传》, 其中 3 842 个不重复汉字, 覆盖古籍常用字库的 86%; 泛化实验所使用的数据集是古籍医书著作《神农本草经》。采用这两种数据集既验证模型在常规语境的能力(《水浒传》), 又检验其在专业领域的泛化性(《神农本草经》)。使用 Python 语言对水浒传文本语料库文本进行清洗, 其中包括去除不必要的 HTML 标签、emoji、数学符号、特殊符号, 以及标点符号规范化、保留古籍必要标点、删除现代标点。对文本进行去重, 使用重复段落检测滑动窗口, 设定窗口大小为 150 个字符。对每个窗口内的 150 个字符进行记录, 比较不同窗口的片段, 如果两个片段相似度达到 60%, 则认为片段重复, 删除其中一个, 从而提升数据质量, 得到一个高质量的文本句子语料库, 得到的语料库按照 8:2 划分为训练集、测试集两部分。

在数据预处理环节, 采用随机截取、随机插入、随机删除和随机交换的文本数据增强和样本均衡处理方法。数据增强中设定的增强比例为 45%, 其中随机截取采用 20% 样本, 截取长度占原句的 25%~45%; 13% 的样本用于随机插入, 句首、句中和句末各插入 33%; 随机删除使用 12% 的样本, 删除 1~3 字, 避开句首和句尾, 避开头尾 3 字符的语义保护机制, 开头一般会包括人名、地名和时间, 删除可能会导致主语缺失, 结尾一般有动作、情感, 删除可能会破坏情节; 10% 的样本用于随机交换, 相邻字进行交换, 每句交

换 2 次。通过动态重采样对样本作均衡处理, 频次小于 15 的字, 其所在句子复制 3 倍, 对于高频字随机丢弃 30% 的样本。

本文实验采用的硬件配置为 Intel(R) Xeon(R) CPU, 2080Ti×2GPU, 系统为 WIN10, Python 版本为 3.7, 配置环境 pytorch=1.8.1, transformers=4.4.1。在训练过程中, Epochs 设置为 120, Batch size 设置为 8, 采用交叉熵损失 (Cross Entropy Loss, CELoss), 使用 Adamw 优化器, 学习率调整策略使用 CosineAnnealingLR。

3.2 评价指标

本文使用的评价指标是 Top-k 准确率, 如式 (20) 所示。模型对每个待预测位置会输出所有可能词的概率分布, Top-k 准确率指模型在所有样本中, 正确的目标词出现在预测的 Top-k 列表中的比例, 该评价指标关注模型是否将正确答案涵盖在预测的前 k 个候选词中。对于一个缺字预测任务中, 往往模型的第一选择并不总是正确的, 可能 在第 3 个、第 5 个, 甚至第 10 个才会出现正确的预测结果。Top-k 准确率越高, 表示模型在前 k 个预测结果中越容易包含正确的预测字词, 即模型性能越好。本文使用 Top-1、Top-3、Top-5、Top-10 对语言模型进行评估。

$$Top_k = (1/N) * \sum (s = 1 to N) [y_s into Top_k(y_s)] \quad (20)$$

其中, N 表示样本数量, y_s 为第 s 个样本的真实标签。

4 实验与结果分析

4.1 模型参数选取实验

为确定中文古籍文本修复模型中 Transformer 编码器的多头注意力机制头数 h 与 Bi-LSTM 的隐藏单元数 v, 本文设计并实施了一系列系统性对照实验。其中, 参数范围设定为 $h \in \{4, 8\}$, $v \in \{128, 256, 512\}$ 。表 1 实验结果显示, 当 h=4 时, 无论隐藏单元数 v 取 64、128 还是 256, 模型在验证集上的准确率均显著低于 h=8 的配置。进一步分析表明, 在 h=8 的条件下, 当 v=256 时, 模型性能达到最优, 准确率最高。因此, 综合对比各项配置的性能表现, 最终确定采用 h=8 和 v=256 作为模型最优参数组合, 以实现古籍文本修复任务的最佳效果。

Table 1 Parameter selection experiments

表 1 参数选取实验

h	v	Top-1 准确	Top-3 准确	Top-5 准确	Top-10 准确
		率/%	率/%	率/%	率/%
4	128	60.44	69.61	73.46	77.82
	256	62.12	71.87	75.91	79.13
	512	65.07	76.51	80.86	85.98
	128	69.22	78.03	83.24	87.56
8	256	71.58	81.35	85.15	90.29
	512	70.06	79.89	83.92	88.76

4.2 对比实验

本文使用多种经典语言模型, N-gram 语言模型、LSTM 模型、Bi-LSTM 模型、Transformer 模型、GTP-2 模型、基于

Bert 的中文预训练模型 Bert-base-chinese 以及本文模型 (Ours) 对数据集进行训练, 得到各模型在训练集和测试集上的表现, 如表 2 和表 3 所示。

Table 2 Prediction results of training sets under different language models

表 2 不同语言模型下训练集预测结果 (%)

模型	Top-1 准确 率	Top-3 准确 率	Top-5 准确 率	Top-10 准确 率
N-Gram	75.48	82.93	85.26	88.05
LSTM	88.04	92.82	94.13	95.52
Bi-LSTM	89.79	94.20	95.40	96.68
LSTM-Attention	90.18	94.34	95.43	96.58
Bi-LSTM-Attention	90.29	94.58	95.72	96.93
Transformer	90.95	95.14	96.19	97.32
GPT-2	93.26	96.53	97.14	98.61
Bert-base-Chinese	95.02	97.81	98.34	98.91
Ours	95.97	98.26	98.74	99.18

Table 3 Prediction results of test sets under different language models

表 3 不同语言模型下测试集预测结果 (%)

模型	Top-1 准确率	Top-3 准确率	Top-5 准确率	Top-10 准确率
N-Gram	33.42	44.98	51.32	60.44
LSTM	46.53	57.87	62.98	70.58
Bi-LSTM	50.72	63.19	68.66	76.21
LSTM-Attention	50.62	61.84	66.79	73.90
Bi-LSTM-Attention	52.00	64.12	69.50	77.03
Transformer	54.62	66.68	72.13	79.27
GPT-2	63.97	74.62	79.06	83.8
Bert-base-Chinese	67.75	77.24	81.02	86.51
Ours	71.58	81.35	85.15	90.29

将各模型在训练集和测试集的结果进行对比可以发现, 本文混合模型对文本预测修复效果显著优于其他语言模型。对比基础 Transformer 模型可知, 训练集准确率提升了 5% 左右, 测试集准确率提升了 17% 左右。与预训练模型 Bert-base-Chinese 相比, 测试集准确率高出 4%, 且训练集和测试集准确率均高于其它语言模型。

4.3 消融实验

为了验证各模块对中文名著文本修复的影响, 在实验过程中进行了消融实验。其中, Ours 是本文提出的混合修复模型, 结果如表 4、表 5 所示。

可以看出, 在 Transformer Encode 单独添加 Global Attention 模块, 训练集 Top-1 准确率从 90.95% 提升至

Table 4 Results of ablation experiment on the training set

表 4 消融实验训练集结果 (%)

模型	Top-1 准确率	Top-3 准确率	Top-5 准确率	Top-10 准确率
Transformer Encode	90.95	95.14	96.19	97.32
Transformer Encoder- Global Attention	92.89	96.10	96.92	97.77
Transformer Encoder- Bi-LSTM	95.21	97.72	98.30	98.88
Ours	95.97	98.26	98.74	99.18

Table 5 Results of ablation experiment on the test set

表5 消融实验测试集结果 (%)

模型	Top-1 准确率	Top-3 准确率	Top-5 准确率	Top-10 准确率
Transformer Encode	54.62	66.68	72.13	79.27
Transformer Encoder- Global Attention	61.08	71.75	76.38	82.74
Transformer Encoder- Bi-LSTM	67.40	77.35	81.53	86.81
Ours	71.58	81.35	85.15	90.29

92.98%, 测试集 Top-1 准确率从 54.62% 提升至 61.08%。这表明了全局注意力有助于捕获序列中的全局依赖关系, 提高了关键特征的表达能力。单独添加 Bi-LSTM 模块后, 训练集 Top-1 准确率从 90.95% 提升至 95.21%, 测试集 Top-1 准确率从 54.62% 提升至 67.40%, 说明了在名著文本修复中, Bi-LSTM 能够捕捉序列的局部时序模式, 尤其是在处理长序列或有时间依赖的任务中。两个模块一起添加后, 训练集 Top-1 准确率从 90.95% 提升至 95.97%, 测试集 Top-1 准确率从 54.62% 提升至 71.58%。消融实验结果表明, 在 Transformer 的基础结合这两模块后, 模型性能优势表现得更加明显。

4.4 泛化性实验

为验证面向破损古籍的全局性语义感知混合修复模型泛化能力, 本文设计了跨数据集实验。选取一部结构和内容风格差异较大的医书古籍《神农本草经》作为泛化性实验数据集, 实验结果如表6所示。

Table 6 Generalization experiments

表6 泛化性实验结果 (%)

模型	Top-1 准确率	Top-3 准确率	Top-5 准确率	Top-10 准确率
Transformer	37.41	49.38	51.15	53.87
Ours	39.91	53.71	55.63	57.56

在模型未进行任何调整的情况下, 模型在《神农本草经》上的修复性能依然较为优异, 说明该方法具备良好的泛化能力。虽然测试集的准确率 Top-1 只达到了 37.82%, 但该方法在文本修复效果中的效果符合预期。实验结果表明, 本文方法不仅在数据集《水浒传》上表现卓越, 对风格和内容均有较大差异的古籍文本《神农本草经》也具有较强的泛化能力, 进一步验证了该模型在古籍文本修复任务中的实用性与鲁棒性。

4.5 模型预测实验

在模型最终模型测试阶段, 用 [mask] 标记语料中所缺失的字, 向模型输入一条包含一个或多个 [mask] 标记的测试集语料, 使用训练好的模型根据上下文信息输出对 [mask] 标记的位置进行预测。模型对随机节选的5条《水浒传》测试语料的输出结果如表7所示, 将模型预测好的语句与原语句进行比对可以看出, 本文方法能够有效地提取名著特有的文学结构特征和丰富的语言风格, 准确地预测出缺失字词, 形成语义连贯、内容完整的句子。模型对《神农本草经》中节选的5条测试语料的输出结果如表8所示。结果表明, 在模型未进行任何调整的情况下, 该方法在文本修复效果中的效果符合预期, 该模型泛化能力较强。

Table 7 Example of text prediction in Water Margin

表7 《水浒传》文本预测示例

缺失语句	模型预测语句	原语句
开封[mask]公文只叫解活的去	开封[府]公文只叫解活的去	开封[府]公文只叫解活的去
宋江进到里[mask]凳子上坐了	宋江进到里[面]凳子上坐了	宋江进到里[面]凳子上坐了
丛林中如[mask]安着得此等之人	丛林中如[何]安着得此等之人	丛林中如[何]安着得此等之人
只见[mask]远远地一个汉子[mask]着一付担桶	只见[了]远远地一个汉子[挑]着一付担桶	只见[远]远远地一个汉子[挑]着一付担桶
后来没半年之[mask], 直抬举高[mask]做到殿帅府太[mask]职事	后来没半年之[间], 直抬举高[俸]做到殿帅府太[尉]职事	后来没半年之[间], 直抬举高[俸]做到殿帅府太[尉]职事

Table 8 Text prediction example of Shennong's Materia Medica

表8 《神农本草》文本预测示例

缺失语句	模型预测语句	原语句
主治五脏积[mask], 胃胀闭	主治五脏积[热], 胃胀闭	主治五脏积[热], 胃胀闭
久食轻身不[mask], 延年神[mask]	久食轻身不[老], 延年神[仙]	久食轻身不[老], 延年神[仙]
痈肿[mask]瘤, 以疮药	痈肿[痛]瘤, 以疮药	痈肿[疮]瘤, 以疮药
病在[mask]腹以下者, 先服药而后[mask]	病在[肠]腹以下者, 先服药而后[食]	病在[心]腹以下者, 先服药而后[食]
主治心腹欬逆邪[mask], 补不[mask], 女子风寒在子[mask], 绝孕[mask]年无子	主治心腹欬逆邪[气], 补不[足], 女子风寒在子[肠], 绝孕[十]年无子	主治心腹欬逆邪[气], 补不[足], 女子风寒在子[宫], 绝孕十年无子

5 结语

本文提出了一种面向破损古籍的全局性语义感知混合修复模型, 该模型有效的结合了 Transformer Encoder、Bi-LSTM 和 Global Attention 模块的优点, 为恢复古籍文本

原貌、保留原文风格和语义等任务提供了一种新的解决方法。该方法在修复准确性、语义一致性和风格保留方面表现优异。但由于训练数据量不够大、文本风格种类不全等原因, 因而对于一些较为复杂的句型结构, 预测出来的字的准确率相对不高。

未来研究中, 可结合外部专家知识库进行修复, 训练

多种风格的古籍文本。同时,考虑文化背景和历史语境等方面,以进一步提升模型修复效果。

参考文献:

- [1] ZHANG Q Y. Research on the mining and utilization of library ancient literature resources based on the perspective of value chain[J]. Library Journal, 2018, 40(10): 83-86.
张青云. 基于价值链视角的图书馆古籍文献资源挖掘与利用研究[J]. 图书馆学刊, 2018, 40(10): 83-86.
- [2] WU G J. Some problems and countermeasures for the protection of ancient books[J]. South China Metal, 2023(4): 60-62.
吴桂娇. 古籍保护的若干问题与应对措施刍议[J]. 南方金属, 2023(4): 60-62.
- [3] PAN J Q. Ancient book restoration: awakening history[J]. Cultural Industry, 2025(1): 26-28.
潘佳岐. 古籍修复唤醒历史[J]. 文化产业, 2025(1): 26-28.
- [4] HOU Y Q Y, QUAN J C, WANG H W. A review of the development of deep learning[J]. Ship Electronic Engineering, 2017, 37(4): 5-9, 111.
侯宇青阳, 全吉成, 王宏伟. 深度学习发展综述[J]. 舰船电子工程, 2017, 37(4): 5-9, 111.
- [5] GAO J H. The practice of integrating traditional and modern technology in the protection of cultural relics in Jinxiang County Museum[J]. Cultural Relics Identification and Appreciation, 2025(2): 34-37.
高金宏. 金乡县博物馆在文物保护中的传统与现代技术融合实践[J]. 文物鉴定与鉴赏, 2025(2): 34-37.
- [6] BROWN P F. Class-based n -gram models of natural language[J]. Computational Linguistics, 1992, 18(4): 467-479.
- [7] LU Y B, LIU C. Research on Skip-ngram model to solve data sparse problem[J]. Automation Technology and Application, 2015, 34(3): 35-37, 46.
鲁一冰, 刘驰. Skip-ngram模型解决数据稀疏问题的研究[J]. 自动化技术与应用, 2015, 34(3): 35-37, 46.
- [8] ELMAN J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2): 179-211.
- [9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [10] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.
- [11] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.
- [12] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pre-training approach[DB/OL]. <https://arxiv.org/abs/1907.11692>.
- [13] CAO X N, WANG J H, YUE K, et al. Folk literature text restoration based on knowledge distillation[J]. Engineering Science and Technology, 2025, 57(6): 119-130.
曹熊能, 王笏辉, 岳昆, 等. 基于知识蒸馏的民间文学文本修复[J]. 工程科学与技术, 2025, 57(6): 119-130.
- [14] SHENG W, LU Y J, LIU W, et al. Research on the restoration of missing texts in ancient Chinese medicine books based on deep learning[J]. Chinese Journal of Library and Information Science and Technology, 2022, 31(8): 1-7.
盛威, 卢彦杰, 刘伟, 等. 基于深度学习的中医古籍缺失文本修复研究[J]. 中华医学图书情报杂志, 2022, 31(8): 1-7.
- [15] TIAN W J. Research on text filling algorithm based on deep learning[D]. Guilin: Guilin University of Electronic Science and Technology, 2021.
田文靖. 基于深度学习的文本填充算法研究[D]. 桂林: 桂林电子科技大学, 2021.
- [16] DUAN Y, LONG H, QU Y Q, et al. Research on irregular interference repair algorithm for text image based on partial convolution[J]. Computer Engineering and Science, 2021, 43(9): 1634-1644.
段炎, 龙华, 瞿于荃, 等. 基于部分卷积的文字图像不规则干扰修复算法研究[J]. 计算机工程与科学, 2021, 43(9): 1634-1644.
- [17] LIU D B, WANG H Q, WANG K, et al. Blind restoration method of missing and sparse text images based on content style transfer[J]. Laser & Optoelectronics Progress, 2022, 59(24): 106-117.
刘栋斌, 王慧琴, 王可, 等. 基于内容风格迁移的残缺稀疏文字图像盲修复方法[J]. 激光与光电子学进展, 2022, 59(24): 106-117.
- [18] ZHENG S T. Research on text filling method and application based on deep learning[D]. Guilin: Guilin University of Electronic Technology, 2021.
郑淑婷. 基于深度学习的文本填充方法及应用研究[D]. 桂林: 桂林电子科技大学, 2021.
- [19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[DB/OL]. <https://arxiv.org/abs/1706.03762>.
- [20] PAN M M, ZHAO Q L. Multi-self-supervised feature fusion method based on feature regression[J/OL]. Computer Engineering, 2025-05-13. <https://doi.org/10.19678/j.issn.1000-3428.0252038>.
潘敏敏, 赵其鲁. 基于特征回归的多自监督特征融合方法[J/OL]. 计算机工程, 2025-05-13. <https://doi.org/10.19678/j.issn.1000-3428.0252038>.
- [21] DENG X. Sentiment analysis of fusion attention mechanism and Bi-LSTM[J]. Science and Technology Innovation, 2025(9): 93-96.
邓翔. 融合注意力机制和Bi-LSTM的情感分析[J]. 科学技术创新, 2025(9): 93-96.
- [22] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[DB/OL]. <https://arxiv.org/abs/1409.0473>.

(责任编辑:孙娟)