计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.009591]

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

基于 BERT 古文预训练模型的实体关系联合抽取①

李智杰, 杨盛杰, 李昌华, 张　颉, 董　玮, 介　军

(西安建筑科技大学 信息与控制工程学院, 西安 710055)
通信作者: 杨盛杰, E-mail: 275598284@qq.com

摘　要: 古汉语文本承载着丰富的历史和文化信息, 对这类文本进行实体关系抽取研究并构建相关知识图谱对于文化传承具有重要作用. 针对古汉语文本中存在大量生僻汉字、语义模糊和复义等问题, 提出了一种基于 BERT 古文预训练模型的实体关系联合抽取模型 (entity relation joint extraction model based on BERT-ancient-Chinese pre-trained model, JEBAC). 首先, 通过融合 BiLSTM 神经网络和注意力机制的 BERT 古文预训练模型 (BERT-ancient-Chinese pre-trained model integrated BiLSTM neural network and attention mechanism, BACBA), 识别出句中所有的 subject 实体和 object 实体, 为关系和 object 实体联合抽取提供依据. 接下来, 将 subject 实体的归一化编码向量与整个句子的嵌入向量相加, 以更好地理解句中 subject 实体的语义特征; 最后, 结合带有 subject 实体特征的句子向量和 object 实体的提示信息, 通过 BACBA 实现句中关系和 object 实体的联合抽取, 从而得到句中所有的三元组信息 (subject 实体, 关系, object 实体). 在中文实体关系抽取 DuIE2.0 数据集和 CCKS 2021 的文言文实体关系抽取 C-CLUE 小样本数据集上, 与现有的方法进行了性能比较. 实验结果表明, 该方法在抽取性能上更加有效, *F*1 值分别可达 79.2% 和 55.5%.

关键词: 古汉语文本; 实体关系抽取; BERT 古文预训练模型; BiLSTM; 注意力; 三元组信息

引用格式: 李智杰,杨盛杰,李昌华,张颉,董玮,介军.基于 BERT 古文预训练模型的实体关系联合抽取.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9591.html

Joint Entity Relation Extraction Based on BERT-ancient-Chinese Pre-trained Model

LI Zhi-Jie, YANG Sheng-Jie, LI Chang-Hua, ZHANG Jie, DONG Wei, JIE Jun

(School of Information and Control Engineering, Xi'an University of Architectural Science and Technology, Xi'an 710055, China)

Abstract: Ancient Chinese texts are rich in historical and cultural information. Studying entity relationship extraction of such texts and constructing related knowledge graphs play an important role in cultural inheritance. Given the large number of rare Chinese characters, semantic fuzziness, and ambiguity in ancient Chinese texts, the entity relation joint extraction model based on the BERT-ancient-Chinese pre-trained model (JEBAC) is proposed. First of all, BERT-ancient-Chinese pre-trained model integrates the BiLSTM neural network and attention mechanism (BACBA), identifies all subject and object entities in sentences, and provides a basis for joint extraction of relation and object entities. Next, the normalized coding vector of the subject entity is added to the embedding vector of the whole sentence to better understand the semantic features of the subject entity in the sentence. Finally, combined with the sentence vector with the characteristics of the subject entity and the prompt information of the object entity, the relationship and object entity in the sentence are jointly extracted by BACBA to obtain all triple information (subject entity, relationship, and object entity) in the sentence. The performance of Chinese entity relation extraction DuIE2.0 datasets and the classical Chinese entity

① 基金项目: 国家自然科学基金 (51878536); 陕西省住房城乡建设科技计划基金 (2020-K09); 陕西省教育厅协同创新中心基金 (23JY038)
收稿时间: 2024-01-18; 修改时间: 2024-02-26; 采用时间: 2024-04-01; csa 在线出版时间: 2024-07-03

relation extraction C-CLUE small sample datasets of CCKS 2021 are compared with that of the existing methods. Experimental results show that the proposed method is more effective in extraction performance, with *F*1 values up to 79.2% and 55.5%, respectively.

Key words: ancient Chinese text; entity relation extraction; BERT-ancient-Chinese pre-trained model; BiLSTM; attention; triple information

古汉语文献作为中华民族宝贵的遗产, 承载着丰富的文化和珍贵的知识. 对这些文献进行深入的分析, 挖掘其中的文本语义信息对于理解古代社会的历史演变和传统文化的发展起着至关重要的作用. 其中, 实体关系抽取作为关键任务, 有助于构建古籍知识图谱, 进一步挖掘其中隐含的文化信息. 然而, 由于古汉语的独特性, 处理古汉语文本面临着许多挑战和困难. 首先, 古汉语中存在大量生僻的汉字和词语, 给实体识别和关系抽取带来了困难. 其次, 古汉语中普遍存在着复义现象, 相同的词语在不同上下文中可能具有不同的含义, 这增加了关系抽取的复杂性. 此外, 古汉语文本通常简洁且富含内涵, 语义信息被压缩在有限的字词中, 进一步增加了实体关系抽取的难度. 为了应对这些挑战, 许多学者致力于开发适用于古汉语文献的分词和词性标注模型. 朱晓等[1]、王晓玉等[2]、程宁等[3]和俞敬松等[4]详细探讨了在“先秦”“明史”等古代汉语文献中, 基于机器学习和深度学习的模型在文本分词和词性标注方面的应用. Wang 等[5]深入探索了古汉语文本的分词和词性标注, 并借鉴于 2021 年发布的面向古文智能处理任务的预训练模型 SikuBERT 和 SikuRoBERTa, 通过进一步优化训练, 开发出了 BERT-ancient-Chinese 预训练模型 (BERT-ancient-Chinese pre-trained model, BACM), 对古汉语文本的智能处理做出了重要贡献.

目前, 对古汉语文本的实体关系抽取研究还没有得到足够的重视, 并且相关的数据集也相对匮乏. 为了弥补这一不足, 韩立帆等[6]介绍了一项基于众包标注系统构建的文言文语言理解测评基准及其数据集. 以“二十四史”作为语料库, 公开了 CCKS 2021 的文言文实体识别与关系抽取的开源数据集 C-CLUE. 在此背景下, 本文提出了一种基于 BERT 古文预训练模型的实体关系联合抽取模型 (JEBAC). 本文的主要工作如下.

(1) 针对古汉语文本中存在大量生僻汉字、语义模糊和复义等问题, 本文构建了 JEBAC 模型.

(2) 为了更加准确地理解句子的语义信息, 本文将 BACM 的输出与 BiLSTM 神经网络和注意力机制相融合, 进行深度特征提取, 以获得更加细致的句子语义特征.

(3) 结合带有 subject 实体语义特征的句子向量和 object 实体的提示信息, 通过 BACBA 实现句中关系和 object 实体的联合抽取, 从而得到句中所有的三元组信息.

1 相关工作

实体及其关系抽取是构建古汉语文本大型知识图谱的关键步骤. 目前, 主要有两大类方法[7]: 流水线抽取方法和联合抽取方法.

1.1 流水线抽取方法

流水线抽取方法将实体识别和关系抽取作为两个独立的子任务进行处理. 首先, 通过各种技术和算法对文本进行实体识别, 确定文本中存在的实体. 然后, 基于实体识别的结果, 使用不同的方法来抽取实体之间的关系. 在流水线方法中, 传统的基于统计的方法过度地依赖人工特征, 并且执行起来非常地耗时费力. 因此, 研究者们提出了基于神经网络的流水线抽取方法. Zeng 等[8]运用卷积神经网络技术, 对文本中词和句子的嵌入向量表示进行特征提取, 然后使用 Softmax 分类器对实体之间关系进行分类, 提升了关系抽取的性能. Xu 等[9]提出了一种新的用于关系分类的神经网络 SDP-LSTM, 利用具有长短期记忆单元的多通道递归神经网络沿着两个实体之间的最短依赖路径 (SDP), 对句子中两个实体之间的关系进行分类, 有效地提高了关系抽取的准确率. Nayak 等[10]提出了一种 multi-factor 的注意力机制, 对长距离的实体进行多依存路径的关系抽取, 有效地提升了长距离实体之间的关系抽取性能. 尽管基于神经网络的流水线方法效果显著, 但这类方法通常面临错误传播和误差积累等问题, 并且忽略了实体和关系之间的相关性.

1.2 联合抽取方法

为了解决流水线方法的实体关系抽取模型存在的问题, 并改善其性能, 学者们提出了基于深度学习的联合抽取方法. Gupta 等[11]构建了 TF-MTRNN 模型, 运用表格填充的思想来解决实体识别和关系抽取, 有效地提升了关系抽取的性能. Katiyar 等[12]在改进 RNN 的基础上, 引入了注意力机制, 提出了一种新型的递归神经网络, 用于实体关系联合抽取. Zheng 等[13]设计了一种将实体关系联合抽取任务转化为序列标注的新方案, 有效地缓解了冗余关系的判断, 进一步提升了实体关系抽取的准确率. Fu 等[14]提出了基于图卷积网络(GCN) 的实体关系抽取模型, 以端到端的方式进行构建, 采用 GCN 的关系加权方法来联合学习实体和关系的表征. 这些方法虽然在解决手动特征工程问题方面取得了显著成效, 但在应对重叠三元组问题上表现不佳, 容易导致三元组的冗余抽取. 针对这个问题, Wei 等[15]构建了 CasRel 层叠式二元标记模型, 将实体关系三元组抽取任务分解为 subject 实体抽取和 relation-object 抽取两个子任务, 显著地提升了关系抽取的性能. 然而, 该模型存在一些不足之处: 首先, 在训练时, subject 实体的选择存在随机性, 没有统一抽取所有的 subject 实体, 这可能导致模型的不稳定性; 其次, 文本的信息没有被充分利用, 导致解码过程出现信息丢失; 此外, 在实体类别不均衡的数据集中, 该模型的性能不佳. 为了改善这些问题, Yu 等[16]将实体关系联合抽取任务分解为 HE 抽取和 TER 抽取两个子任务, 并采用 span 标记将它们细化为多个序列标记任务. 通过这种方法, 可以更好地捕捉不同步骤之间的语义依赖性, 从而提高实体识别和关系抽取的交互性.

根据文献 [15,16]的研究基础, 本文构建了 JEBAC 模型. JEBAC 中的 BACM, 能够有效地理解古汉语文本的语义特征. 为了进一步深入古汉语文本的特征提取, 将 BACM 和 BiLSTM 神经网络以及注意力机制相融合, 增强 JEBAC 对文本信息的利用效果. 此外, 通过将 subject 实体的头尾位置信息编码归一化加权融入到句子中, 能够有效地将 subject 实体的特征向量表示添加到句子嵌入向量中. 为了提升对 object 实体的识别, 还引入了辅助识别 object 实体的部分, 以帮助 JEBAC 抽取到更准确的 object 实体. 最后, 采用适用于关系和 object 实体联合抽取的二进制分类器, 有效地识别 subject 实体相关关系下的 object 实体, 并抽取句中所有的三元组.

2 JEBAC 模型

JEBAC 模型的整体架构如图 1 所示, 主要包括: 编码模块、实体识别模块、关系和 object 实体联合抽取模块. JEBAC 模型利用 BACM 获取句子的文本特征, 以得到包含上下文信息的句子嵌入向量表示$\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n]$. 然后, 通过将句子向量输入到 BiLSTM 神经网络和 Self-Attention 机制中, 进行深度特征提取, 以获得更加准确的句子语义特征表示$\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n]$. 在持续迭代学习和训练的过程中, 采用二进制序列标注的方法, 利用 subject 实体和 object 实体的二进制分类器, 识别出句中的所有 subject 实体和 object 实体. 同时, 将 subject 实体的头、尾位置信息编码向量 $\boldsymbol{v}_s^1$、$\boldsymbol{v}_s^2$ 归一化加权融入到句子向量中, 得到带有 subject 实体特征表示的句子嵌入向量$\boldsymbol{E}'$. 通过关系和 object 实体联合抽取的二进制分类器, 抽取出句中所有的实体关系三元组 (s, r, o), 其中 s 表示 subject 实体, r 表示 relation 关系, o 表示 object 实体.

2.1 编码模块

BACM 句子嵌入编码: 由于 BERT 预训练模型能够快速有效地理解句子文本中携带的特征信息, 且允许使用大规模训练数据进行下游任务的再训练. 在 BERT fine-tuned 的基础上, BACM 通过对大规模古汉语文本数据进行训练而产生. JEBAC 模型通过使用 BACM 提取句子序列的语义特征, 得到输入文本的句子嵌入向量. 给定序列长度为 n 的第 m 个输入文本的句子表示, 如式 (1):

$$\boldsymbol{S}_m=[w_{m_1},w_{m_2},\cdots,w_{m_n}],\ m=1,2,\cdots,batch \tag{1}$$

其中, 第 i 个字w_{m_i}的向量表示, 如式 (2):

$$\boldsymbol{e}_i=\boldsymbol{E}_{\text{token}}(w_{m_i})+\boldsymbol{E}_{\text{seg}}(\boldsymbol{S}_m)+\boldsymbol{E}_{\text{pos}}(i) \tag{2}$$

其中, $\boldsymbol{E}_{\text{token}}$、$\boldsymbol{E}_{\text{seg}}$、$\boldsymbol{E}_{\text{pos}}$ 分别表示词嵌入、句子嵌入和位置嵌入. 输入句子 $\boldsymbol{S}_m$ 经过 BACM 编码后, 输出具有上下文语义特征表示的句子嵌入向量 $\boldsymbol{E}$, 如式 (3) 所示:

$$\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n]=\text{BACM}(\boldsymbol{S}_m) \tag{3}$$

BACBA 对句子文本的语义特征加强编码: BiLSTM 神经网络是一种独特的循环神经网络, 它具备逆向编码的能力, 可以通过利用句子中后面的重要信息, 有效地得到句子的双向语义依赖, 对输入的句子向量深层次化特征表示. 具体操作: 将 $\boldsymbol{E}$ 输入到 BiLSTM 神经网

络中进行编码, 在每个 t 时刻的输入, 不仅包括字向量, 还包括 t–1 时刻的输出向量. 每个 t 时刻的输出 $\boldsymbol{h}_t$ 由前、后向编码向量拼接而成. 具体过程表示如式 (4):

$$\boldsymbol{h}_t = \text{BiLSTM}(\boldsymbol{E}_t, \boldsymbol{h}_{t-1}) \tag{4}$$

其中, $\boldsymbol{E}_t$ 表示在 t 时刻输入的字向量. 经过 BiLSTM 神经网络编码之后, 输出为$\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n]$.

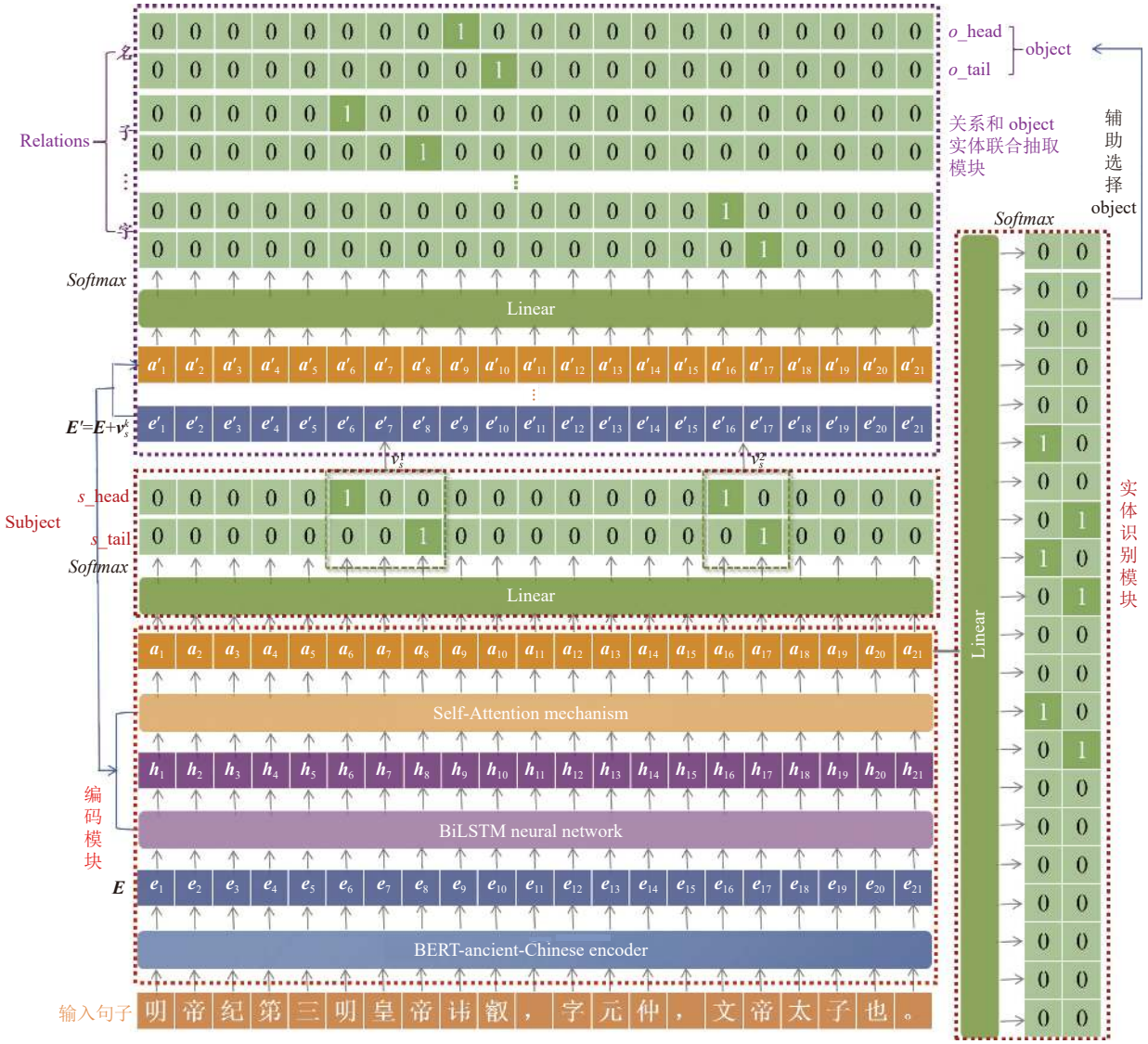

图 1　JEBAC 模型的架构

为了加强深度特征编码, 增强 JEBAC 模型对文本的解读, 在 BiLSTM 神经网络中加入 Self-Attention 机制, 为特征 $\boldsymbol{H}$ 每个部分的输入赋予不同的权重, 深度提取关键信息, 使得 JEBAC 能够充分地利用文本的特征. 如式 (5) 所示:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

其中, $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 是输入的字向量矩阵, 分别表示查询矩阵、键矩阵和值矩阵; $\sqrt{d_k}$ 是 $\boldsymbol{K}$ 矩阵第 1 个维度的平方根, 它能够保持梯度的平衡. $\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$ 表示输入句子中的每个字向量之间的关系计算过程. 经过 Self-Attention 机制后, 得到输出$\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n]$.

2.2　实体识别模块

JEBAC 模型的实体识别模块是 s 实体和 o 实体识别的解码部分. 具体采用 0/1 二进制分类器来识别 s 实体和 o 实体的头 (head) 和尾 (tail) 的位置. 计算公式如下:

$$p_i^{s\_\text{head}} = Sigmoid(\boldsymbol{W}_{\text{head}}^{s}\boldsymbol{A}_i + \boldsymbol{b}_{\text{head}}^{s}) \tag{6}$$

$$p_i^{s\_\text{tail}} = Sigmoid(\boldsymbol{W}_{\text{tail}}^{s}\boldsymbol{A}_i + \boldsymbol{b}_{\text{tail}}^{s}) \quad (7)$$

$$p_i^{o\_\text{help\_head}} = Sigmoid(\boldsymbol{W}_{\text{head}}^{o}\boldsymbol{A}_i + \boldsymbol{b}_{\text{head}}^{o}) \quad (8)$$

$$p_i^{o\_\text{help\_tail}} = Sigmoid(\boldsymbol{W}_{\text{tail}}^{o}\boldsymbol{A}_i + \boldsymbol{b}_{\text{tail}}^{o}) \quad (9)$$

其中, $\boldsymbol{A}_i$是输入句子中第 i 个字向量, $p_i^{s_\text{head}}$和$p_i^{s_\text{tail}}$表示 s 实体的第 i 个字向量经解码处理后输出的头、尾位置概率值. 若该值大于实验设置的阈值 0.5, 则该字所在位置将被标记为 1, 否则将被标记为 0. $p_i^{o_\text{help_head}}$和$p_i^{o_\text{help_tail}}$同理. $\boldsymbol{W}_*^*$和$\boldsymbol{b}_*^*$是训练权重和偏置, *Sigmoid* 表示相应的激活函数.

2.3 关系和 object 实体联合抽取模块

JEBAC 模型的关系和 object 实体联合抽取模块结合带有 s 实体特征的句子嵌入向量编码和 o 实体的提示信息, 针对所有的 r 进行 o 实体的解码. 具体采用级联多层 0/1 二进制分类器来识别 s 条件下的 r 和 o. 这部分需要预定义当前数据集所要提取的关系, 关系的数量即为二进制标注的层数. 它的输入是加入了 s 实体特征表示的句子嵌入向量$\boldsymbol{E}'$, 计算表示如式 (10):

$$\boldsymbol{E}' = \boldsymbol{E} + \boldsymbol{v}_s^k \quad (10)$$

其中, $\boldsymbol{v}_s^k$是每个 s 实体头、尾位置归一化编码的向量表示. 对$\boldsymbol{E}'$进行解码时, 对于所有的 r, 标注指针将同时为每个识别的 s 实体标记出相应的 o 实体. 具体过程如下:

$$p_i^{o\_\text{head}} = Sigmoid(\boldsymbol{W}_{\text{head}}^{r\text{-}o}\boldsymbol{E}_i' + \boldsymbol{b}_{\text{head}}^{r\text{-}o}) \quad (11)$$

$$p_i^{o\_\text{tail}} = Sigmoid(\boldsymbol{W}_{\text{tail}}^{r\text{-}o}\boldsymbol{E}_i' + \boldsymbol{b}_{\text{tail}}^{r\text{-}o}) \quad (12)$$

其中, $\boldsymbol{E}'$ 是第 i 个字的向量表示, $p_i^{o_\text{head}}$和$p_i^{o_\text{tail}}$表示 o 实体的第 i 个字向量经解码处理后输出的头、尾位置概率值, 阈值同样设置为 0.5. $\boldsymbol{W}_*^*$和$\boldsymbol{b}_*^*$是训练权重和偏置, *Sigmoid* 表示相应的激活函数.

2.4 损失函数

JEBAC 模型的损失函数是 s 实体头尾预测的损失、o 实体头尾预测的损失和 r 及 o 实体头尾联合预测的损失 3 个部分的加和. 这 3 个部分都采用二进制分类器预测, 使用二分类交叉熵损失函数 (binary cross entropy loss, BCELoss). 具体计算过程如式 (13):

$$Loss = \sum_{e\in E} -\frac{1}{n}\sum_{i=1}^{n}(y_i^e \cdot lbp_i^e + (1-y_i^e)\cdot lb(1-p_i^e)) \quad (13)$$

其中, E = {*s*_head, *s*_tail, *o*_help_head, *o*_help_tail, *o*_head, *o*_tail}, n 为输入句子的长度. y_i^e表示句子中第 i 个字是 e 情况的真实样本值 (是/否, 0/1), p_i^e表示 e 情况的二进制分类器预测结果的概率.

3 实验

3.1 数据集与评价指标

JEBAC 模型中的 BACM 部分是基于 BERT fine-tuned, 且它的词汇表中包括 BERT 的词汇表. 所以, 本文的实验部分选择中文实体关系抽取 DuIE2.0 数据集, 以验证 JEBAC 模型的可信度. 在 CCKS 2021 文言文实体关系抽取 C-CLUE 小样本数据集上, 验证 JEBAC 模型在古汉语文本抽取中的优越性.

DuIE2.0 数据集[17]: 百度提供的业界规模最大的基于 schema 的中文关系抽取数据集. 该数据集包含超过 43 万三元组、21 万中文句子和 48 个预定义的关系类型.

C-CLUE 数据集[6]: CCKS 2021 文言文实体识别与关系抽开源数据集. 它是一个基于众包标注系统构建的文言文语言理解测评基准及数据集, 由天津大学数据库课题组贡献, 使用“二十四史”作为语料库进行标注. “二十四史”是中国古代各朝撰写的二十四部史书的总称, 记录了丰富的历史人物和事件. 该数据集包含近 5 000 个三元组和 25 个预定义的关系类型. 数据集统计见表 1.

表 1 数据集统计

数据集	训练集	测试集
DuIE2.0	171 135	20 652
C-CLUE	2 550	712

实验使用准确率 P、召回率 R 和 $F1$ 值作为评价指标, 计算公式如式 (14)–式 (16):

$$P = \frac{\text{所有句子中预测正确的三元组个数}}{\text{所有句子中预测出的三元组个数}} \times 100\% \quad (14)$$

$$R = \frac{\text{所有句子中预测正确的三元组个数}}{\text{所有句子中人工标注的三元组个数}} \times 100\% \quad (15)$$

$$F1 = \frac{2\times P\times R}{P+R} \times 100\% \quad (16)$$

C-CLUE 数据集的格式和关系表: C-CLUE 数据集以 JSON 格式存储, 包括句子文本及其相应的三元组, 如下所示.

{
“text”: “王益怒, 遣人告枢密使蒋玄晖与何太后私通, 杀玄晖而焚之, 遂弑太后于积善宫. 又杀宰相柳璨, 太常卿张延范车裂以徇”,

“spo_list”: [{
“predicate”: “任职”,
“object_type”: “JOB”,
“subject_type”: “PER”,
“object”: “宰相”,
“subject”: “柳璨”
}, {
“predicate”: “任职”,
“object_type”: “JOB”,
“subject_type”: “PER”,
“object”: “太常卿”,
“subject”: “张延范”
}, {
“predicate”: “杀”,
“object_type”: “PER”,
“subject_type”: “PER”,
“object”: “张延范”,
“subject”: “王”
}, {
“predicate”: “杀”,
“object_type”: “PER”,
“subject_type”: “PER”,
“object”: “柳璨”,
“subject”: “王”
}]
}

对应的关系表以 csv 格式存储, 如表 2 所示.

表 2 C-CLUE 数据集关系表

关系	映射值	关系	映射值
属于	0	朋友	13
葬于	1	弟	14
讨伐	2	位于	15
兄	3	名	16
号	4	归属	17
字	5	出生地	18
任职	6	作	19
同名于	7	作战	20
隶属于	8	子	21
姓	9	父	22
管理	10	杀	23
升迁	11	去往	24
依附	12		

3.2 实验环境与参数设置

实验在 Windows 10 下进行, 使用的硬件配置包括 Intel® Core(TM) i7-8700 CPU 处理器和 NVIDIA GeForce RTX 2080Ti 显卡. 编程语言使用 Python 3.7, 模型的优化器选择 AdamW, 模型的训练采取早停机制的策略进行. 模型的参数设置如表 3.

表 3 模型参数设置

参数	DuIE2.0	C-CLUE
Batch_size	16	6
Embedding_dim	768	768
Hidden_dim	256	256
Learning_rate	1E–5	1E–5
BiLSTM_layers	2	2
Epoch	200	200
All_threshold	0.5	0.5
L2_regularization	0.01	0.01
Patience	8	8

3.3 实验结果与分析

本文旨在联合抽取古汉语文本中实体和关系的三元组信息, 以便构建相关的知识图谱. 因此, 通过与目前较好的实体关系联合抽取模型进行对比实验, 验证 JEBAC 模型的有效性.

MHS[18]: 是一种实体关系联合抽取模型, 利用 CRF 层, 将实体识别和关系抽取任务全部转化为多目标选取问题.

CasRel[15]: 是一种端到端的层叠式标记模型, 成功地解决了句中三元组重叠问题, 主要包括 subject 标记模块和 relation-object 标记模块. 该模型将关系建模成 subject 映射到 object 的函数, 可以同时抽取多个实体关系三元组.

CopyMTL[19]: 是 CopyRE 模型的优化, 采用 BiLSTM 神经网络在编码阶段获取句子的全局特征, 进而创建句子的特征向量. 在解码阶段, 通过融合 Attention-LSTM 对关系、头实体和尾实体进行逐个的预测和识别.

WDec[20]: 是一种解决实体关系联合抽取中三元组重叠问题的 encoder-decoder 模型.

JEBAC 模型与上述模型的各项性能指标比较结果如表 4.

表 4 不同模型实验结果 (%)

模型	DuIE2.0			C-CLUE		
	P	*R*	*F*1	*P*	*R*	*F*1
CopyMTL	49.9	39.4	43.9	35.3	29.1	31.9
WDec	64.1	54.2	58.7	39.8	54.4	45.9
MHS	77.2	62.3	69.0	35.1	10.0	15.6
CasRel	77.2	80.1	78.6	41.5	49.1	45.0
JEBAC	78.1	80.3	79.2	45.0	72.3	55.5

可以看出, JEBAC 模型在 DuIE2.0 数据集上的 *F*1 值比 CasRel 模型提高了 0.6%, *P* 值提高了约 1%, *R* 值基本持平. 这说明了 JEBAC 模型预测的三元组个

数比 CasRel 模型预测的少, 预测正确的三元组个数没有很大提升, 即 JEBAC 模型有效地降低了预测三元组的冗余. 通过对比, JEBAC 模型在 C-CLUE 数据集上效果显著, 比 CasRel 模型的 *F*1 值高了 10.5%, 比数据源[7]的关系抽取任务 *F*1 值 47.61%, 提高了约 7.9%, 这证明了 JEBAC 模型在古汉语文本数据集上的有效性.

JEBAC 模型在 C-CLUE 数据集中的每个预定义关系下的抽取结果, 如表 5.

表 5 在 C-CLUE 数据集上的关系抽取结果 (%)

关系	*P*	*R*	*F*1
属于	18.4	50.0	26.9
葬于	83.3	55.6	66.7
讨伐	58.8	100.0	74.0
兄	36.4	52.2	42.9
号	48.6	65.4	55.7
字	0.0	0.0	0.0
任职	50.1	70.0	58.4
同名于	38.9	67.7	49.4
隶属于	40.9	90.0	56.3
姓	0.0	0.0	0.0
管理	45.5	50.0	47.6
升迁	70.0	58.3	63.6
依附	50.0	61.5	55.2
朋友	0.0	0.0	0.0
弟	25.5	52.2	34.3
位于	60.4	87.9	71.6
名	52.8	73.7	61.5
归属	42.9	85.7	57.1
出生地	47.7	100.0	64.6
作	31.1	100.0	47.5
作战	27.3	60.0	37.5
子	44.1	80.0	56.8
父	96.7	7.7	14.3
杀	48.4	100.0	65.2
去往	48.8	83.0	61.4

表 5 的结果说明了小样本数据集 C-CLUE 的缺陷. 由于它在某些关系下的样本数据集过于少, 导致 JEBAC 模型对于这些关系的学习性能不佳, 抽取不到这些关系. 在 C-CLUE 数据集上, JEBAC 模型的稳定性较差.

3.4 消融实验分析

为了验证 JEBAC 模型基于 BACM 对文言文数据集的效果, 以及 object 辅助识别模块和 BiLSTM+Att 模块对 JEBAC 模型的效果, 在 C-CLUE 数据集上进一步完成了消融实验, 实验结果如表 6 所示.

表 6 中, –object 表示对 JEBAC 模型减少 object 辅助识别模块, –BA 表示对 JEBAC 模型减少 BiLSTM+Att 模块. 对比前两行的数据, JEBAC 模型基于 BACM 比基于 BERT 的效果更好, 能够更有效地表示古汉语文本的特征. JEBAC 模型在消去 object 辅助识别模块时, *P* 值提高了 1.2%, *R* 值降低了 7.6%, 说明预测三元组的冗余个数变多了, 影响了模型的性能. JEBAC 模型在消去 BA 模块后, *P* 和 *R* 明显降低, 这里体现了 BiLSTM 神经网络和注意力机制的重要性. BA 模块能够使 JEBAC 模型学习到句子中每个字的深度特征, 加强相关特征向量表示, 从而提高模型抽取的性能.

表 6 在 C-CLUE 数据集上的消融实验结果 (%)

模型	*P*	*R*	*F*1
JEBAC	45.0	72.3	55.5
$\text{JEBAC}_{\text{BERT}}$	44.9	63.0	52.4
$\text{JEBAC}_{\text{–object}}$	46.2	64.7	53.9
$\text{JEBAC}_{\text{–BA}}$	42.1	64.6	51.0

图 2 和图 3 展示了这些模型随着训练周期 Epoch 的增长, 损失值 Loss 和 *F*1 值的变化对比. 在训练过程中使用了早停机制, BA 模块的融合会使模型的训练周期变长. BACM 融合 BA 模块和 object 辅助识别模块共同作用, 提高了 JEBAC 模型对古汉语文本实体关系抽取的性能.

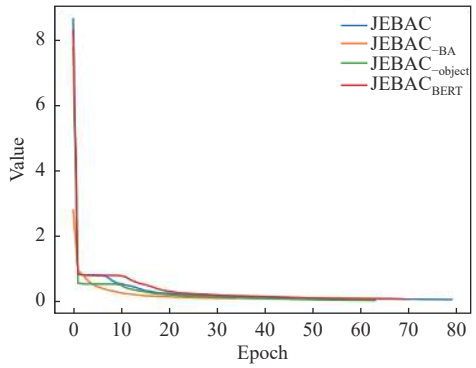

图 2 模型训练的 Loss

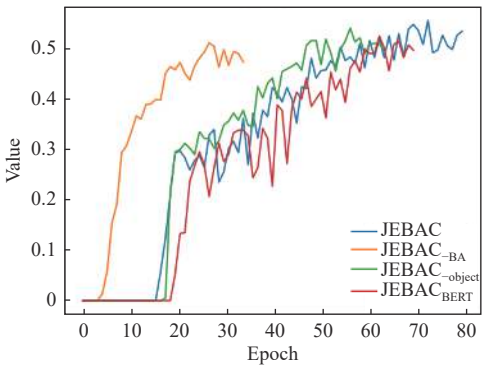

图 3 模型测试的 *F*1

3.5 JEBAC 模型的预测

本文随机选取《史记》中的《史记卷二·夏本纪》《史记卷六·秦始皇本纪》和《史记卷七·项羽本纪》这 3 个文本进行预测, 验证 JEBAC 模型的有效性. 预测结果如表 7.

表 7 JEBAC 模型的预测结果

古汉语文本	句子	预测三元组
《史记卷二·夏本纪》	235	24
《史记卷六·秦始皇本纪》	863	94
《史记卷七·项羽本纪》	531	161

表 7 中, 《史记卷二·夏本纪》文本共有 235 个句子, 使用 JEBAC 模型共识别出了与约束关系相关的 24 个三元组: {[(“黄帝”, “子”, “禹”)], [(“鯀”, “子”, “禹”), (“昌意”, “任职”, “臣”), (“鯀”, “任职”, “臣”)], [(“鯀”, “去往”, “羽山”)], [(“舜”, “子”, “禹”), (“鯀”, “子”, “禹”)], [(“伯禹”, “任职”, “司空”)], [(“鯀”, “子”, “禹”)], [(“蒙、羽”, “出生地”, “徐州”)], [(“阳鸟”, “出生地”, “淮海”)], [(“土、梦”, “出生地”, “九江”)], [(“黑水”, “出生地”, “梁州”)], [(“泾”, “出生地”, “西河”)], [(“织皮”, “出生地”, “西戎”)], [(“夔”, “任职”, “乐”)], [(“舜”, “子”, “商均”)], [(“皋陶”, “位于”, “许”)], [(“帝禹”, “去往”, “会稽”)], [(“帝禹”, “子”, “启”)], [(“帝中康”, “同名于”, “中康”)], [(“帝扃”, “子”, “帝廑”)], [(“帝孔甲”, “同名于”, “孔甲”)], [(“孔甲”, “子”, “帝皋”)]}. 这些三元组中有预测错误的需要人工检查, 然后将预测正确的结果保留, 从而构建《史记卷二·夏本纪》的知识图谱.

4 结论

本文提出了一种基于 BERT 古文预训练模型的实体关系联合抽取模型 (JEBAC). 通过将 BACM 输出的句子嵌入向量融入 BiLSTM 神经网络和 Self-Attention 机制中, 进行深度特征提取, 以获得更加准确的句子语义特征的向量表示. 同时, 结合带有 subject 实体语义特征表示的句子编码向量和提示的 object 实体信息, 联合抽取对应的关系和 object 实体, 从而抽取句中所有的三元组信息. 实验结果表明, 本文的 JEBAC 模型在 DuIE2.0 数据集上抽取性能有所提升, 在小样本文言文数据集 C-CLUE 上效果显著. 尽管 JEBAC 模型在性能方面得到了一定程度的改进, 但其稳定性仍有待提高. 构建大规模古汉语文本数据集是进一步提高 JEBAC 模型性能和稳定性的解决办法. 此外, JEBAC 模型对于如何抽取 zero 三元组是一个挑战.

目前, 本文研究的 JEBAC 模型主要针对公共数据集进行测试. 接下来的工作将深入落实到大规模古汉语文本数据中, 并对该模型进行相应的改进和参数优化, 为相关领域的知识图谱构建及应用提供坚实的基础.

参考文献

1 朱晓, 金力. 条件随机场图模型在《明史》词性标注研究中的应用效果探索. 复旦学报 (自然科学版), 2014, 53(3): 297–304. [doi: 10.15943/j.cnki.fdxb-jns.2014.03.001]

2 王晓玉, 李斌. 基于 CRFs 和词典信息的中古汉语自动分词. 数据分析与知识发现, 2017, 1(5): 62–70.

3 程宁, 李斌, 葛四嘉, 等. 基于 BiLSTM-CRF 的古汉语自动断句与词法分析一体化研究. 中文信息学报, 2020, 34(4): 1–9. [doi: 10.3969/j.issn.1003-0077.2020.04.001]

4 俞敬松, 魏一, 张永伟, 等. 基于非参数贝叶斯模型和深度学习的古文分词研究. 中文信息学报, 2020, 34(6): 1–8. [doi: 10.3969/j.issn.1003-0077.2020.06.002]

5 Wang PY, Ren ZC. The uncertainty-based retrieval framework for ancient Chinese CWS and POS. Proceedings of the 2nd Workshop on Language Technologies for Historical and Ancient Languages. Marseille: European Language Resources Association, 2022. 164–168.

6 韩立帆, 季紫荆, 陈子睿, 等. 数字人文视域下面向历史古籍的信息抽取方法研究. 大数据, 2022, 8(6): 26–39. [doi: 10.11959/j.issn.2096-0271.2022058]

7 张仰森, 刘帅康, 刘洋, 等. 基于深度学习的实体关系联合抽取研究综述. 电子学报, 2023, 51(4): 1093–1116. [doi: 10.12263/DZXB.20221176]

8 Zeng DJ, Liu K, Lai SW, *et al*. Relation classification via convolutional deep neural network. Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. Dublin: ACL, 2014. 2335–2344.

9 Xu Y, Mou LL, Li G, *et al*. Classifying relations via long short term memory networks along shortest dependency paths. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 1785–1794. [doi: 10.18653/v1/D15-1206]

10 Nayak T, Ng HT. Effective attention modeling for neural relation extraction. Proceedings of the 23rd Conference on Computational Natural Language Learning. Hong Kong: ACL, 2019. 603–612.

11 Gupta P, Schütze H, Andrassy B. Table filling multi-task recurrent neural network for joint entity and relation

extraction. Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016. 2537–2547.

12 Katiyar A, Cardie C. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: ACL, 2017. 917–928.

13 Zheng SC, Wang F, Bao HY, *et al*. Joint extraction of entities and relations based on a novel tagging scheme. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017. 1227–1236.

14 Fu TJ, Li PH, Ma WY. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 1409–1418.

15 Wei ZP, Su JL, Wang Y, *et al*. A novel cascade binary tagging framework for relational triple extraction. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 1476–1488.

16 Yu BW, Zhang ZY, Shu XB, *et al*. Joint extraction of entities and relations based on a novel decomposition strategy. Proceedings of the 24th European Conference on Artificial Intelligence: ECAI 2020. Santiago de Compostela: IOS Press, 2020. 2282–2289.

17 Li SJ, He W, Shi YB, *et al*. DuIE: A large-scale Chinese dataset for information extraction. Proceedings of the 8th CCF International Conference on Natural Language Processing and Chinese Computing. Dunhuang: Springer, 2019. 791–800.

18 Bekoulis G, Deleu J, Demeester T, *et al*. Joint entity recognition and relation extraction as a multi-head selection problem. Expert Systems with Applications, 2018, 114: 34–45. [doi: 10.1016/j.eswa.2018.07.032]

19 Zeng DJ, Zhang HR, Liu QY. CopyMTL: Copy mechanism for joint extraction of entities and relations with multi-task learning. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 9507–9514. [doi: 10.1609/aaai.v34i05.6495]

20 Nayak T, Ng HT. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 8528–8535. [doi: 10.1609/aaai.v34i05.6374]

(校对责编: 孙君艳)