文章编号:1008-1542(2025)02-0141-10

面向中医医案的事件抽取方法研究

马月坤[1,2,3,4],崔漠晓[1]

(1. 华北理工大学人工智能学院,河北唐山　063210;
2. 河北省工业智能感知重点实验室,河北唐山　063210;
3. 北京科技大学计算机与通信工程学院,北京　100083;
4. 材料领域知识工程北京市重点实验室,北京　100083)

摘　要:针对中医事件抽取过程中存在事件论元边界模糊的问题,综合运用卷积神经网络、双向长短期记忆网络和注意力机制提出了一种局部与全局语义特征融合的事件抽取模型(event extraction model integrating local and global semantic feature,EE-LGSF),以提升中医事件抽取的效果。首先,通过结合不同滤波窗口大小的卷积神经网络提取文本的多维度局部特征信息,同时利用双向长短期记忆网络捕捉文本的全局特征信息;其次,通过门控机制实现局部与全局信息的动态交互,以增强对论元边界的识别能力;再次,引入模糊跨度注意力机制,动态调整注意力范围,从而优化论元跨度的决策过程;最后,以条件随机场进行标签预测。结果表明,所提模型比相关模型在中医医案数据集上的 F1 值提升了 3.0～11.0 个百分点,在解决中医事件抽取问题方面表现更为优异。所提模型能有效利用文本局部和全局语义信息,提高对文本跨度学习的灵活性来增强模型对论元边界的识别能力,从而获得更好的中医事件抽取效果,对中医知识传承和发展具有参考价值。

关键词:自然语言处理;事件抽取;中医医案;注意力机制;卷积神经网络;动态融合;跨度

中图分类号:TP391.1;R241　　**文献标识码**:A　　**DOI**:10.7535/hbkd.2025yx02003

Research on event extraction methods for medical records of traditional Chinese medicine

MA Yuekun[1,2,3,4], CUI Moxiao[1]

(1. College of Artificial Intelligence, North China University of Science and Technology, Tangshan, Hebei 063210, China;
2. Hebei Provincial Key Laboratory of Industrial Intelligent Perception, Tangshan, Hebei 063210, China;
3. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China;
4. Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China)

Abstract: In response to the issue of fuzzy event argument boundaries in traditional Chinese medicine (TCM) event extraction, an event extraction model integrating local and global semantic features (EE-LGSF) was proposed, which combined

收稿日期:2024-12-09;修回日期:2025-02-17;责任编辑:胡姝洋
基金项目:国家重点研发计划项目(2022YFC3502303);河北省工业智能感知重点实验室项目(SZX2021013)
第一作者简介:马月坤(1976—),女,河北廊坊人,教授,博士,主要从事自然语言处理、知识工程方面的研究。
E-mail:mayuekun@163.com
马月坤,崔漠晓.面向中医医案的事件抽取方法研究[J].河北科技大学学报,2025,46(2):141-150.
MA Yuekun, CUI Moxiao. Research on event extraction methods for medical records of traditional Chinese medicine[J]. Journal of Hebei University of Science and Technology, 2025, 46(2): 141-150.

convolutional neural networks, bidirectional long short-term memory networks, and attention mechanisms to enhance the effectiveness of TCM event extraction. Firstly, multi-dimensional local feature information of the text was extracted by combining convolutional neural networks with different filter window sizes, while the global feature information of the text was captured using bidirectional long short-term memory networks. Secondly, on this basis, dynamic interaction between local and global information was achieved through gating mechanisms to enhance the ability of model to identify argument boundaries. Furthermore, a fuzzy span attention mechanism was introduced to dynamically adjust the attention range, thereby optimizing the decision-making process for argument spans. Finally, label prediction was performed using conditional random fields. The results indicate that the proposed model improves the F1 score by 3.0 to 11.0 percentage points on the TCM medical records data-set, demonstrating superior performance in addressing TCM event extraction issues compared to related models. The proposed model effectively leverages both local and global semantic information of the text, enhances the flexibility of span learning and improves the capability of the model to identify argument boundaries, thereby achieving better performance in TCM event extraction. It has reference value for the inheritance and development of TCM knowledge.

Keywords: natural language processing; event extraction; traditional Chinese medicine medical records; attention mechanism; convolutional neural network; dynamic fusion; span

中医医案是直观反映中医诊疗过程的文字描述形式,这些医案中含有大量的医学诊疗知识,通过信息抽取,将这些知识转化为结构化形式,可以更好地为医生提供辅助诊疗。以往中医医案的信息抽取工作主要聚焦于对症状、疾病、方剂等静态知识及其关系的抽取,但是这种抽取方式已经无法满足对更广泛学习、推理和理解中医诊疗场景的需求。而事件是更加细化、动态的知识表示形式,具有更深入和丰富的语义表示能力[1],能更好地理解和把握事物发展的本质和规律。中医的实际诊疗过程是一个多层级、多角度的复杂体系,是中医医生依据先验中医知识,结合疾病的具体症状模式、病因、病性和病位等对疾病进行指导治疗的过程[2],可以视为由多个环环相扣的事件构成。

当前在中医领域,文献[3-5]分别通过模式匹配的方式和基于 BERT-CRF 的深度学习方法进行事件抽取,但由于中医事件论元的文本表达具有内聚性且上下文语义关联较弱,导致出现论元边界模糊的情况,这一现象对中医事件抽取任务造成了挑战。当前,在针对论元边界模糊问题的研究中,文献[6]通过先识别事件论元跨度再回归调整边界的方式精确定位事件论元;文献[7]提出结合正反向解码器的方法提升事件论元识别准确性;文献[8]融合 ERNIE 和 MacBERT 的语义特征优化模型性能,有效解决了论元边界模糊问题。然而以上研究都缺少了对文本局部特征的学习,导致模型会忽略不同字符间的语义关联。此外,在以往的研究中,对论元跨度的学习过于依赖给定标注数据中的跨度边界[9-10],这样会使得模型缺乏对于跨度边界识别的灵活性,导致识别效果降低。

因此,本文采取一种局部和全局特征融合的方式,使用不同尺寸卷积核的卷积神经网络(convolutional neural networks,CNN)通过串联和并联结合的方式提取不同尺度下的文本局部特征,同时使用双向长短期记忆网络(bidirectional long short-term memory,BiLSTM)实现对全局语义特征的提取,最后使用模糊跨度注意力机制(fuzzy span attention,FSA),通过线性偏差学习位置嵌入,并引入掩码函数改变不同距离字符间的注意力权重,使得模型更聚焦于文本的特定部分,以解决中医事件抽取中存在的论元边界模糊的问题。

1 局部与全局语义特征融合的事件抽取模型

局部与全局语义特征融合的事件抽取模型(event extraction model integrating local and global semantic feature,EE-LGSF)的整体结构如图 1 所示。模型主要包括局部特征提取模块、全局特征提取模块、特征融合模块和解码模块。模型的具体流程如下:对于给定的一条文本 $X=\{x_1,x_2,\cdots,x_N\}$,模型首先使用 BERT 对文本进行词嵌入得到对应的向量表示 $\boldsymbol{E}=\{e_1,e_2,\cdots,e_N\}$。接着,将 BERT 输出的同时输入局部特征提取模块和全局特征提取模块。局部特征提取模块采用双分支结构,每个分支包含 4 个卷积通道,通过不同尺寸的卷积核提取多尺度局部特征,捕捉文本的局部上下文信息。全局特征提取模块则建模文本的全局语义依赖关系,获取全局语义特征。随后,特征融合模块通过动态加权策略融合局部和全局特征,并结合模糊跨度注意力机制增强对论元边界的感知。最后,解码模块基于融合后的特征预测每个字符的标签。

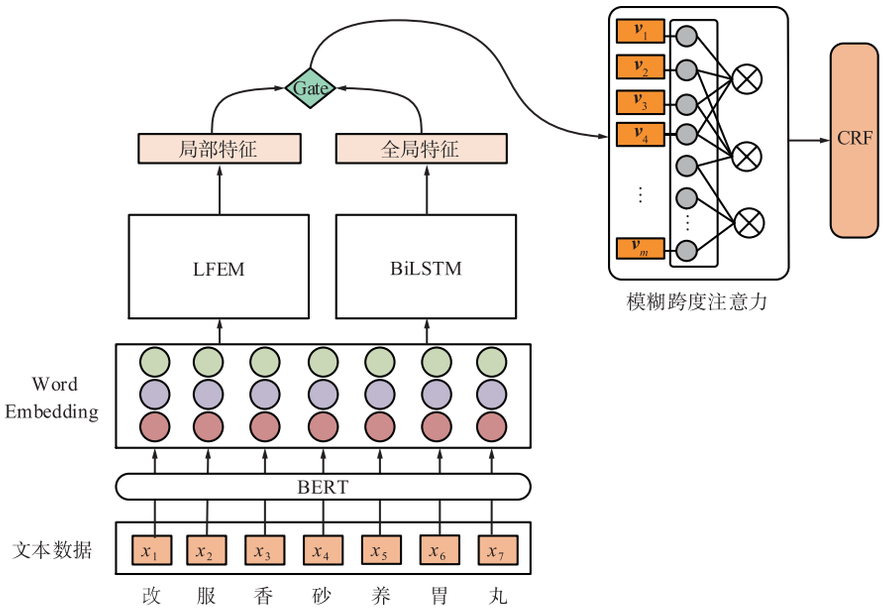

图 1　EE-LGSF 模型结构

Fig. 1　EE-LGSF model structure

1.1　局部特征提取模块

为了解决中医论元边界模糊的问题，本文提出了一个局部特征提取模块（local feature extraction module，LFEM），在 BERT 模型提供的全局上下文信息指导下，通过卷积神经网络从长序列文本中提取局部信息。具体而言，该方法采用不同尺度的卷积核学习当前字符与前几个字符的关联性，并将前一卷积层的输出组合到下一层，以便捕捉不同字符长度下的序列特征。局部特征提取模块包含 2 个分支，每个分支设有 4 个卷积通道，如图 2 所示。每条通道由卷积层、Relu 激活函数及池化层组成，卷积核尺寸分别为 3、4、5 和 6。

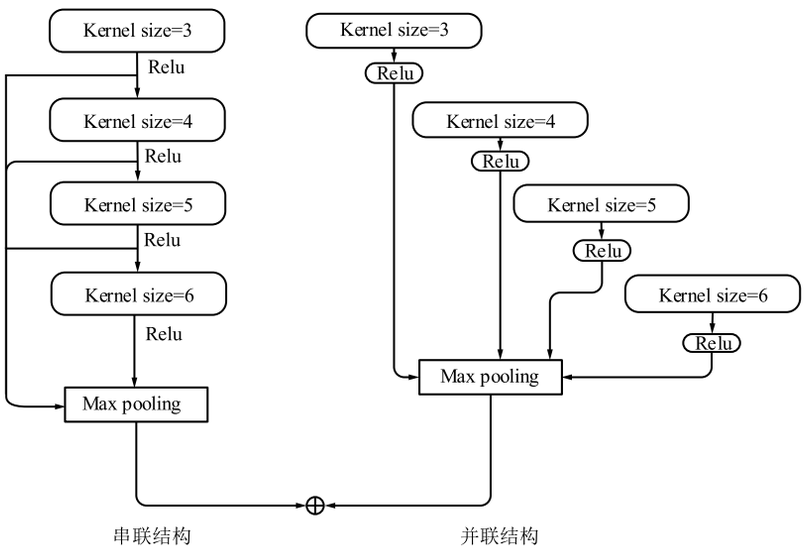

图 2　LFEM 模块结构

Fig. 2　LFEM modular structure

对于串联部分，输入文本首先通过卷积核尺寸为 3 的卷积层提取局部特征，并经过 Relu 激活函数优化反向传播效率，防止梯度问题，如式(1)所示。

$$S_i^1 = \mathrm{Relu}(w \cdot T_{i:i+h-1} + b), \qquad w \in R^{h*k}, \quad i = 3, \tag{1}$$

式中：S_i^1 表示卷积核尺寸为 3 的文本局部特征；w 表示维度为 $h*k$ 的卷积核；h 表示不同大小的滑动窗口

的个数；k 为输入的文本序列中每个词的向量维度；$T_{i:i+h-1}$ 表示输入矩阵的第 i 行到第 $i+h-1$ 行组成的滑动窗口；b 为偏置参数。随后，将生成的文本特征逐步输入到不同尺寸卷积核的卷积网络中。最后，通过最大池化层保留最显著的特征，如式(2)、(3)所示。

$$S_i = \mathrm{Relu}(\mathrm{Conv}(S_{i-1})), \quad w \in R^{h*k}, \quad i = 4,5,6, \tag{2}$$

$$\mathrm{Pooling}(S_i), \quad i = 3,4,5,6。 \tag{3}$$

为了更大范围地获取字符间的语义信息，并保留每个卷积网络在各自尺度下的独有特征，本文将经过首个卷积层处理后的原始向量特征予以保留，并使其与不同卷积层输出的特征并行处理，如图2的并联部分所示，这一策略允许同时处理多个特征层。最终，模型将卷积层串联生成的特征 S_i 经过最大池化层后与卷积层并联生成的特征 Sm_i 经过最大池化层后通过连接操作融合，从而最终得到模型提取的综合文本局部特征Feature。具体的计算如式(4)、(5)所示。

$$Sm_i = \mathrm{Relu}(w \cdot T_{i:i+h-1} + b), \quad w \in R^{h*k}, \quad i = 3,4,5,6, \tag{4}$$

$$\mathrm{Feature} = \mathrm{Pooling}(S_i) \oplus \mathrm{Pooling}(Sm_i) 。 \tag{5}$$

1.2 全局特征提取模块

BiLSTM通过2个LSTM网络实现：一个沿正向顺序处理输入序列，另一个沿反向顺序。在每个时间步骤中，每个LSTM单元都能访问前后文的信息，从而有效捕捉双向语义依赖关系。

这种双向处理能力使BiLSTM在理解和建模序列中的上下文信息方面表现出色，特别是在涉及双向依赖的任务中效果明显。在事件抽取任务中，为了全面学习句子的语义结构，每个单词都需要充分包含上下文信息。因此，本文采用BiLSTM作为文本全局特征提取器。BiLSTM的特征提取过程如式(6)—(8)所示。

$$\vec{h}_t = \overrightarrow{\mathrm{LSTM}}(\vec{h}_{t-1}, e_t) , \tag{6}$$

$$\overleftarrow{h}_t = \overleftarrow{\mathrm{LSTM}}(\overleftarrow{h}_{t-1}, e_t) , \tag{7}$$

$$h_t = [\vec{h}_t ; \overleftarrow{h}_t], \tag{8}$$

式中：$\vec{h}_t$ 为BiLSTM在 t 时刻向右方向的输出；$\overleftarrow{h}_t$ 为BiLSTM在 t 时刻向左方向的输出；$\vec{h}_{t-1}$ 和 $\overleftarrow{h}_{t-1}$ 分别为BiLSTM在 $t-1$ 时刻向2个方向的输出；h_t 为BiLSTM的最终输出。

1.3 特征融合模块

本文采用门控机制(Gate)与模糊跨度注意力机制来对全局和局部特征进行融合。门控机制可以动态选择最相关的特征，有效增强模型对关键信息的记忆和保留，同时抑制不相关的信息。模糊跨度注意力机制则有助于捕捉输入中不同距离元素间的关系，增强模型对上下文的理解。

首先，通过门控机制融合2种特征，如式(9)、(10)所示。门向量 $\boldsymbol{g}$ 由一个带有Sigmoid激活函数的全连接层生成，用以控制特征输入。门向量 $\boldsymbol{g}$ 的范围在0～1之间，表示每个输入特征的重要性。

$$\boldsymbol{g} = \sigma(W_g[\boldsymbol{p};\boldsymbol{q}] + b_g) , \tag{9}$$

$$\mathrm{Gate}(\boldsymbol{p},\boldsymbol{q}) = \boldsymbol{g} \odot \boldsymbol{p} + (1-\boldsymbol{g}) \odot \boldsymbol{q}, \tag{10}$$

式中：$\boldsymbol{p}$、$\boldsymbol{q}$ 为输入向量；$\sigma(\cdot)$为Sigmoid激活函数；$\odot$为逐元素乘积；[;]为拼接操作；W_g、b_g 均为可训练参数。将模块提取的局部特征和全局特征经过门控机制进行动态融合，如式(11)所示，输出结果保留了对模型最有效的特征。

$$\boldsymbol{V} = \mathrm{Gate}(\boldsymbol{F},\boldsymbol{H}) , \tag{11}$$

式中：$\boldsymbol{F}$ 为局部特征提取模块提取的局部特征向量；$\boldsymbol{H}$ 为全局特征提取模块提取的全局特征向量；$\boldsymbol{V}$ 为经过门控机制的最终输出向量。

本文在门控机制后采用了FSA，这种机制通过设定一个模糊范围并根据编码自适应调整跨度长度，来提升对论元边界的跨度感知。模型在学习过程中学习到的跨度不是精确的，而是模糊的。由于中医内存在很多相似的论元，例如“腰腹痛、腰腹冷痛、腰腹酸痛”都是腰腹痛的一种症状表现，但这种细微的差异可能导致对症状的描述产生显著的变化，因此模糊的跨度使得模型不局限于精确的边界，而是灵活地探索多种可能的论元。以“腰腹痛”论元为例，精准边界和模糊边界分别如图3、图4所示。

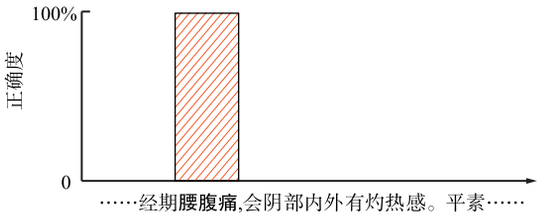

图 3　“腰腹痛”的精准边界

Fig. 3　Precise boundary of "lower back and abdominal pain"

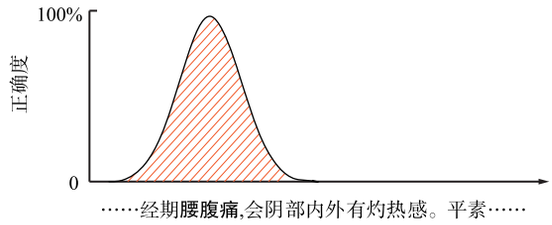

图 4　“腰腹痛”的模糊边界

Fig. 4　Fuzzy boundary of "lower back and abdominal pain"

FSA 是基于相对位置编码(relative positional encoding，RPE)的多头自注意力机制，多头注意力由多个单头注意力组合而成，每个注意力头侧重不同的相关性[11]。每个注意力的输入包括查询向量(query，$\boldsymbol{q}$)、键向量(key，$\boldsymbol{k}$)以及值向量(value，$\boldsymbol{v}$)，其中 $\boldsymbol{q}$ 用于衡量与每个 $\boldsymbol{k}$ 的相关性。$\boldsymbol{q}$ 向量通常与当前位置或整个序列信息相关，$\boldsymbol{k}$ 向量代表的是输入序列中每个位置的信息，可视为每个位置的“索引”或“查找表”。在计算注意力权重时，每个位置的 $\boldsymbol{k}$ 向量会与 $\boldsymbol{q}$ 向量进行比较，通过点积来评估 $\boldsymbol{q}$ 与序列中各位置之间的相关性。

由于位置信息对于识别论元边界起着重要作用，在逐个查询 $\boldsymbol{q}$ 和 $\boldsymbol{k}$ 点积之后，添加了一个静态的、非学习的线性偏置[12]作为相对位置嵌入，这种方法能够降低模型对位置嵌入的“过度拟合”，从而提升模型的性能。计算的可视化视图如图 5 所示。

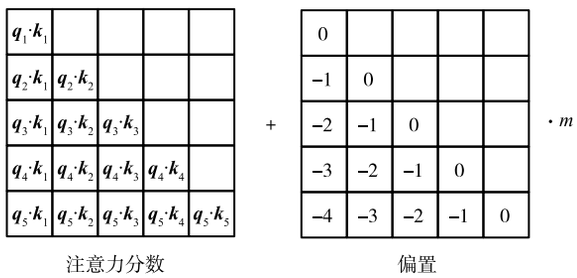

图 5　相似度计算可视化视图

Fig. 5　Similarity matrix calculation visualization

在计算每个头部的注意力分数时，为每个注意力分数添加一个恒定的偏置(见图 5)，该偏置依赖于 $\boldsymbol{q}$ 和 $\boldsymbol{k}$ 之间的距离。因此对于序列中位置 t 处的词元(token)，首先得到每个注意力头计算的该 token t 与序列中的 token r 的相似度。相似度 s_{tr} 可以通过式(12)表示。

$$s_{tr}=\boldsymbol{y}_t\boldsymbol{W}_{\boldsymbol{q}}\boldsymbol{W}_{\boldsymbol{k}}^{\mathrm{T}}\boldsymbol{y}_r^{\mathrm{T}}+m\cdot[-(t-r)]\ ,\tag{12}$$

式中：$\boldsymbol{W}_{\boldsymbol{k}}$ 为 $\boldsymbol{k}$ 的权重矩阵；$\boldsymbol{W}_{\boldsymbol{q}}$ 为 $\boldsymbol{q}$ 的权重矩阵；$\boldsymbol{y}_t$ 为字符 t 的向量表示；$\boldsymbol{y}_r$ 为字符 r 的向量表示；m 表示固定的头部特定斜率，对于注意力头数为 8 的模型，使用的斜率是几何顺序$\frac{1}{2^1}$、$\frac{1}{2^2}$、…、$\frac{1}{2^8}$。

最后，通过 Softmax 函数获得相应的注意力权重 α_{tr}，如式(13)所示。

$$\alpha_{tr}=\frac{\exp(s_{tr})}{\sum_{q=0}^{t-1}\exp(s_{tq})}\ ,\tag{13}$$

式中：exp 为指数函数；q 为序列中某一位置，取值为 $0\sim t-1$。

为了动态调整注意力的范围并控制每个字符的注意力得分，本文引入了 PENG 等[13]提出的掩码函数 g_m。假设可能的注意力跨度的最大长度是 L_{span}，那么新的注意力分数 α'_{tr} 可以表示为式(14)。

$$\alpha'_{tr}=\frac{g_m(t-r)\exp(s_{tr})}{\sum_{q=t-L_{\mathrm{span}}}^{t-1}g_m(t-r)\exp(s_{tq})},\tag{14}$$

式中掩码函数 g_m 由注意力改变函数 g_α 决定，因此首先需要确定模糊跨度的注意力改变函数 g_α，然后基于此构建用于跨度感知表征学习的掩码函数 g_m。在模糊跨度注意力机制中，注意力权重并不是简单地在某个边界之外被完全置为零，而是随着距离的增加而逐渐降低，显示出一种“衰减”模式。这意味着，计算某个位置的注意力权重时，较近的上下文信息会被赋予更高的权重，而较远的信息权重会逐渐减少。因此，模型能够利用一定范围内的上下文信息，同时更关注与当前目标相关的邻近信息。FSA 通过一个可学习的参数 $\delta\in[0,1]$调整注意力跨度的长度。$g_\alpha(z)$和相应的 $g_m(z)$如式(15)—(17)所示。

$$g_a(z)=\frac{-z+l+d}{d}\ ,\tag{15}$$

$$l=\delta L_{\text{span}}\ ,\tag{16}$$

$$g_m(z)=\begin{cases}1, & g_a(z)>1,\\ 0, & g_a(z)<0,\\ g_a(z), & \text{otherwise},\end{cases}\tag{17}$$

式中：d 为控制衰减的注意力范围的长度的超参数；l 为控制全注意力范围的长度；z 为2个词元之间的距离。

图6为函数 g_m 的示意图。其中在 $z\leqslant l$ 的范围内为全注意力范围，此时 g_m 为1；在长度为 d 的范围内，g_m 呈单调递减，表示注意力衰减的范围，此时会根据不同位置编码的距离减少关注程度；在 $z\geqslant(l+d)$ 时 g_m 为0，表明距离过远的 token 便不再受关注。

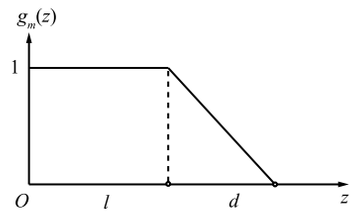

图6 函数 g_m 示意图

Fig. 6 Schematic diagram of function g_m

FSA 通过在不同的头部上独立学习注意力跨度的长度，实现了最佳的模糊跨度调整。FSA 的模糊跨度注意力层是在 Transformer 编码器的基础上新增的，不参与编码过程，因此其最大跨度范围虽受限，但仅影响跨度决策，对序列中标记的表示不产生影响。

1.4 解码模块

条件随机场(confitional random fields，CRF)通过考虑相邻标签之间的依赖关系，有效地得到最优标签序列。首先，输入序列 $X=\{x_1,x_2,\cdots,x_t\}$ 经过特征表示层后，生成隐藏状态上下文特征向量 $\boldsymbol{H}=\{h_1,h_2,\cdots,h_t\}$，这些特征向量用于表示输入序列中每个位置的上下文信息。然后，CRF 通过特征函数和权重计算每个可能的标签序列 y 的分数，如式(18)所示。

$$\text{score}(x,y)=\sum_i\sum_k\lambda_k f_k(x,y,i)\ ,\tag{18}$$

式中：f_k 代表特征函数；k 表示特征函数的个数；i 表示序列中的位置索引；λ_k 表示对应的权重；x 为输入序列；y 为标签序列。之后通过 Softmax 函数，CRF 计算归一化后的标签序列 y 的概率，确保所有标签序列的概率之和为1，如式(19)、(20)所示。

$$P(y\mid x)=\frac{\mathrm{e}^{\text{score}(x,y)}}{Z(x)}\ ,\tag{19}$$

$$Z(x)=\sum_{y'}\mathrm{e}^{\text{score}(x,y')}\ ,\tag{20}$$

式中：score(x,y)为位置分数；$P(y|x)$为归一化概率函数；$Z(x)$为归一化因子。最后，使用 Viterbi 算法通过动态规划找到每个状态的最优路径，确保最大化条件概率。这一过程使 CRF 能够有效捕捉标签之间的依赖关系，从而在序列标注任务中取得优异的效果。

2 实验设计

2.1 数据获取与预处理

本文使用多个中医医案集作为语料来源。选取了《妇科圣手杨宗孟临床56年经验集》[14]、《张淑亭延嗣医案》[15]、《钱伯煊妇科医案》[16] 3种医案。在中医领域，中医事件并不是和传统意义上的事件描述完全一致，而是根据该文本的特点及使用需求分成了具有中医诊疗特色的事件，因此本文根据中医的诊疗过程将中医事件分为了病情、检查、辨证、治疗。病情事件是指涉及患者个人体征的变化及因疾病表现出的异常情况，主要记录在现病史部分，详述症状、症状产生的原因及变化趋势；检查事件主要描述患者进行的各项医疗检查，包括检查内容、检查结果、检查时间；辨证事件是指医家根据四诊(望、闻、问、切)信息分析疾病的证候；治疗事件是指患者病历中记录的与治疗措施和方案相关的内容，包括治法、方剂和其他治疗方法，如手术、针灸等。以《张淑亭延嗣医案》中的医案数据为例，每种事件所对应的论元及论元角色如表1所示。

表 1　事件类型对应的论元与论元角色

Tab. 1　Argument and argument role corresponding to the event type

事件类别	实例	事件论元角色	事件论元
病情	出现小腹痛，傍晚身热，大量黄白色脓样白带从阴道而出，臭气难闻。胸闷纳差，头昏心烦，溲赤便干，舌质红，苔黄厚，脉数	症状	小腹痛、傍晚身热、大量黄白色脓样白带从阴道而出、臭气难闻、胸闷纳差、头昏心烦、溲赤便干
		舌象	舌质红
		苔象	苔黄厚
		脉象	脉数
检查	6 月 25 日经行之际，做诊断性刮宫术，病理报告：晚期分泌期宫内膜，分泌欠佳	检查结果	晚期分泌期宫内膜、分泌欠佳
		检查时间	6 月 25 日经行之际
		检查内容	诊断性刮宫术
辨证	证属湿热之毒，损伤任带	证候	湿热之毒、损伤任带
治疗	治宜清热解毒，利湿止带。拟五味消毒饮合止带方化裁	治法	清热解毒、利湿止带
		方剂	五味消毒饮合止带方

本文将医案数据转化为文本并进行数据清洗，例如文本数据里含有的空格、额外增添的字符等，这些无关噪声会影响实验的效果，因此需要对数据进行清洗。对于在文本中出现的符号、日期，以及检查结果中出现的数字也进行规范化处理，之后根据不同事件类型对医案进行筛选并划分为不同长度的事件句，在不影响原始文本语义的情况下将长句进行切分。在切分过程中，保留原文本的格式，并使用逗号、分号、句号等符号对文本进行切分。最终将数据集按 8∶1∶1 比例分为训练集、测试集和验证集。具体数据见表 2。

表 2　实验数据集

Tab. 2　Experimental dataset

数据集	句子数量/句	字符数量/个	实体数量/个
训练集	3 872	76 527	4 002
测试集	484	9 670	420
验证集	484	9 528	411

2.2　事件标注

本文选用三位序列标注法 BIO(B-begin，I-inside，O-outside)，B-X 表示论元 X 的起始位置，I-X 表示论元 X 的中间位置，O 表示不是论元。事件抽取中事件类型的识别依赖于不同的事件触发词，但由于中医医案文本具有口语化特点，触发词可能缺失或难以识别。因此，本文采用了事件类型与论元角色统一标注的策略，例如在辨证事件“证属肾阳虚衰，带下不孕，带脉失约，任督失职”中，“肾阳虚衰”作为辨证事件的论元，标注为肾(B-BZSJ-ZH)阳(I-BZSJ-ZH)虚(I-BZSJ-ZH)衰(I-BZSJ-ZH)。另一个例子，病情事件“平素带下清稀，淋漓不止，味腥难闻，腰酸肢冷”中的“腰酸肢冷”标注为腰(B-BQSJ-ZZ)酸(I-BQSJ-ZZ)肢(I-BQSJ-ZZ)冷(I-BQSJ-ZZ)。各事件类型的论元角色及其对应标签如表 3、表 4 所示。

表 3　事件标注方法

Tab. 3　Event annotation method

事件类别	论元角色	首字	中间
病情事件	症状	B-BQSJ-ZZ	I-BQSJ-ZZ
	舌象	B-BQSJ-SX	I-BQSJ-SX
	苔象	B-BQSJ-TX	I-BQSJ-TX
	脉象	B-BQSJ-MX	I-BQSJ-MX
检查事件	检查内容	B-JCSJ-JCNR	I-JCSJ-JCNR
	检查结果	B-JCSJ-JCJG	I-JCSJ-JCJG
	检查时间	B-JCSJ-JCSJ	I-JCSJ-JCSJ
辨证事件	证候	B-BZSJ-ZH	I-BZSJ-ZH
治疗事件	治法	B-ZLSJ-ZF	I-ZLSJ-ZF
	方剂	B-ZLSJ-FJ	I-ZLSJ-FJ
	中药	B-ZLSJ-ZY	I-ZLSJ-ZY

表 4　BIO 标注示例

Tab. 4　Example of BIO annotation

论元	标注
证	O
属	O
肾	B-BZSJ-ZH
阳	I-BZSJ-ZH
虚	I-BZSJ-ZH
衰	I-BZSJ-ZH
，	O
带	B-BZSJ-ZH
下	I-BZSJ-ZH
不	I-BZSJ-ZH
孕	I-BZSJ-ZH

2.3　实验评价指标及参数设置

本实验中，BERT 隐藏状态维度为 768，学习率设置为 2×10^{-5}，模型中其他参数的学习率为 1×10^{-4}。BiLSTM 的隐藏状态维度为 384，BiLSTM 的隐藏层层数为 2，Dropout 设置为 0.5，Epoch 为 20，注意力头数为 8，优化器采用 Adam。

本文使用事件抽取任务中最常用的评价指标：精确率(P)、召回率(R)、F1 值来评估模型性能。

2.4 对比模型

目前使用不同的预训练语言模型是在特定领域上事件抽取的常用方法，因此本文选用BERT-CRF[17]、ALBERT-CRF[18]、RoBERTa-CRF[19]作为一组对照模型。BERT-CRF结合了BERT的特性和CRF的序列标注能力，将事件提取转换为BIO任务，利用CRF确保序列标注的一致性和规则性。ALBERT-CRF利用ALBERT动态词向量进行特征提取。RoBERTa-CRF使用RoBERTa通过在大规模语料库上动态掩码预训练，优化了语言表示能力，能够捕捉复杂的上下文依赖。

另外选取其他3种事件抽取方法作为另一组对照模型：Lattice-LSTM-CRF[20]是一种网格结构的长短时记忆网络中文紧急事件抽取模型；BERT-BiGRU-CRF通过BiGRU处理时序信息，而CRF优化标签序列的预测，确保输出的连贯性；BERT-IDCNN-BiLSTM-CRF[21]融合了BERT的上下文表示、IDCNN的局部特征提取、BiLSTM的长距离依赖捕捉与CRF层的标签一致性，旨在提升文本理解和标注精度。

3 实验结果与分析

3.1 模型对比结果

从表5实验结果可以看出，本文模型在性能上显著优于其他基线方法模型。在第1组对照模型中，本文模型相较于BERT-CRF和RoBERTa-CRF在F1值上分别提升了5.40和3.98个百分点，与ALBERT-CRF模型相比F1值提升了6.01个百分点，这验证了本文模型在捕捉语义特征方面的优势。

表5 实验结果

Tab.5 Experimental result

模型	*P*	*R*	F1
BERT-CRF	70.31	68.70	69.50
ALBERT-CRF	71.54	66.43	68.89
RoBERTa-CRF	74.17	67.97	70.92
Lattice-LSTM-CRF	71.43	58.39	64.26
BERT-BiGRU-CRF	74.80	70.23	72.44
BERT-IDCNN-BiLSTM-CRF	69.49	62.60	65.86
ours	**75.78**	**74.05**	**74.90**

在第2组对照模型中，本文模型相较于BERT-BiGRU-CRF模型在F1值上提升了2.46个百分点，这表明通过引入多尺度卷积核的卷积神经网络，能够更全面地捕捉文本的多维特征，从而显著提升对正样本的预测能力。与BERT-IDCNN-BiLSTM-CRF模型相比，本文模型在F1值上提升了9.04个百分点，凸显了本文模型在捕捉字符间细粒度语义关联方面的优越性。此外，本文模型相较于Lattice-LSTM-CRF模型在F1值上提升了10.64个百分点。这是由于Lattice-LSTM-CRF模型的网格结构虽然能够捕捉词汇间的边界信息，但由于中医论元的内聚性强、上下文语义关联较弱，模型容易将一些非中医论元的词汇错误识别为有效论元，而本文模型通过模糊跨度注意力机制和多尺度的卷积神经网络，能够增强对论元上下文信息的捕捉和利用能力，从而提升事件抽取的效果。总体而言，本文模型基本上取得了最好的性能，验证了该模型的有效性。

3.2 消融实验

为验证本文模型各组件的效果，在中医医案数据集上进行以下消融实验：

1)WCNN 比较了只使用并联卷积层的情况，以验证串联和并联结合的局部特征提取模块的有效性；

2)WFSA 在EE-LGSF模型基础上去除了模糊跨度注意力机制；

3)Gate+ 在EE-LGSF模型基础上，用加法操作代替了门控机制；

4)EE-LGSF.1 去除了门控机制和局部特征提取模块；

5)EE-LGSF.2 使用普通注意力机制替代模糊跨度注意力机制。

表6 消融实验结果

Tab.6 Results of ablation experiment

模型	*P*	*R*	F1
WCNN	71.97	72.52	72.24
WFSA	74.79	67.94	71.20
Gate+	73.77	68.70	71.15
EE-LGSF.1	73.81	70.99	72.37
EE-LGSF.2	73.48	74.05	73.76
ours	**75.78**	**74.05**	**74.90**

消融实验结果如表6所示。可知移除任一模块都会导致F1值下降，证明每个组件都对中医事件抽取任务有积极影响。具体分析如下。

1)WCNN 不使用迭代卷积神经网络时，F1值下降，表明相比仅使用单一卷积核，迭代使用不同尺度的卷积核能更全面地学习不同序列长度下的文本特征，提高识别的准确性。

2)WFSA 去掉模糊跨度注意力机制后，F1值有所下降，表明了模糊跨度注意力对论元边界的判断及

跨度学习的重要性。

3)Gate+　将特征融合方式由门控机制改为加法后，F1 值显著降低，验证了门控机制在筛选全局与局部信息中的有效性。

4)EE-LGSF.1　特征双通道架构在综合全局与局部信息提取中的优势，对比仅依赖全局特征的模型，更能有效捕捉关键信息。

5)EE-LGSF.2　相比于使用普通注意力机制，模糊跨度注意力机制通过动态调整覆盖范围，减少了全范围关注带来的冗余，提高了模型的精准度和效率。

综上所述，本文提出的特征提取方法和信息融合策略对中医事件抽取任务具有显著优势。

3.3　不同卷积核参数对实验结果的影响

在中医医案数据集上进行多种实验以探究 EE-LGSF 中不同卷积核尺寸对中医事件抽取效果的影响。实验设置了 3 种卷积核参数配置：

情况 1：卷积核尺寸为 2、3、4、5。

情况 2：卷积核尺寸分别为 3、4、5、6，这是 EE-LGSF 的标准配置。

情况 3：将 LFEM 中串联结构部分的卷积核尺寸改为 2、3、4、5，如图 7 所示。

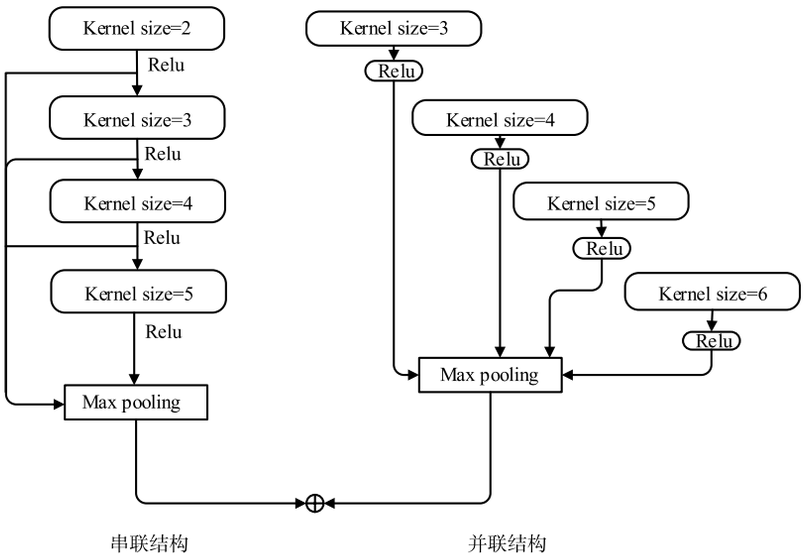

图 7　LFEM 模块更改卷积核参数后的结构

Fig. 7　Structure of LFEM module after changing convolution kernel parameters

实验结果如图 8 所示。当卷积核参数设置为情况 1 时，模型的 F1 值、精确率 P 和召回率 R 均低于其他 2 种情况。这表明情况 1 中的 4 个卷积核尺寸较小，虽然能够捕捉更细腻的局部特征，但缺乏对于中医较长论元的语义捕捉。设置情况 3 时，其效果优于本文模型。这说明，在串联使用卷积层的部分，使用较小尺寸的卷积核更有利于捕捉细粒度特征；而在并联使用卷积层的部分，使用较大尺寸的卷积核可以使模型考虑更大范围的上下文信息，从而增强其语义理解能力。

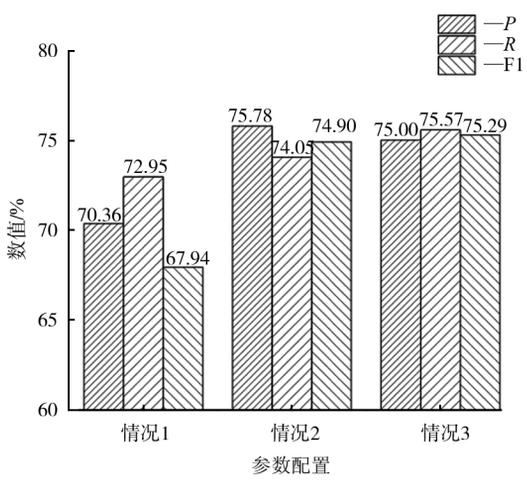

图 8　使用不同卷积核参数的实验结果

Fig. 8　Experimental results using different convolution kernel parameters

4　结　语

提出了一种面向中医医案的事件抽取方法，该方法考虑到中医论元的内聚性较强和上下文语义关联较弱的特征，采用具有多尺度卷积核的卷积神经网络，通过叠加多个不同尺寸大小的卷积核，以串联

和并联的方式捕获多维度的局部特征信息，从而增强中医论元内在的局部依赖关系，同时结合 BiLSTM 进行双通道特征提取，以提升对中医论元边界的感知能力。此外，模型还利用门控机制对局部与全局特征信息进行有效筛选和融合。在 Transformer 编码层中引入模糊跨度注意力机制，调整注意力焦点，进一步提升对论元的上下文语义关联的捕捉能力。本文模型在中医医案文本的事件抽取任务中的表现优于其他对比模型，展示了其在处理中医特有语义结构方面的有效性。

本文仅从中医医案的角度探索了中医的事件抽取方法，然而中医诊疗过程是多种理论的综合体，所以如何从更广泛的中医知识中抽取事件并构建事件图谱，以及如何实现大规模语料标注，是未来研究的重要方向。

参考文献/References：

[1] 胡志磊，靳小龙，陈剑赟，等. 事件图谱的构建、推理与应用[J]. 大数据，2021，7(3)：80-96.
HU Zhilei，JIN Xiaolong，CHEN Jianyun，et al. Construction，reasoning and applications of event graphs[J]. Big Data Research，2021，7(3)：80-96.

[2] 王瑞坤. 面向临床决策支持的《伤寒论》事件知识图谱构建与应用研究[D]. 北京：北京中医药大学，2023.

[3] 仝磊. 面向中医医疗文本的智能化事件发现与分类方法研究[D]. 北京：北京交通大学，2023.
TONG Lei. Research on Intelligent Event Discovery and Classification Methods for Traditional Chinese Medicine Texts[D]. Beijing：Beijing Jiaotong University，2023.

[4] 王乐. 基于深度学习的中医临床病历事件抽取方法研究[D]. 北京：北京交通大学，2021.
WANG Le. Research on Event Extraction Method of TCM Clinical Records Based on Deep Learning[D]. Beijing：Beijing Jiaotong University，2021.

[5] 高甦，陶浒，蒋彦钊，等. 中医文献的句子级联合事件抽取[J]. 情报工程，2021，7(5)：15-29.
GAO Su，TAO Hu，JIANG Yanzhao，et al. Sentence-level joint event extraction of traditional Chinese medical literature[J]. Technology Intelligence Engineering，2021，7(5)：15-29.

[6] 赵宇豪. 基于跨度回归的中文事件抽取方法研究与应用[D]. 贵阳：贵州大学，2023.

[7] 熊枫. 基于角色信息引导和双向跨度检测的篇章级事件论元抽取研究[D]. 武汉：华中师范大学，2023.
XIONG Feng. Research on Document-Level Event Argument Extraction Based on Role Information Guidance and Bidirectional Span Detection[D]. Wuhan：Central China Normal University，2023.

[8] 沙明洋，张思佳，傅庆财，等. 基于动态权重的多模型集成水产动物疾病防治事件抽取方法[J]. 华中农业大学学报，2023，42(3)：80-87.
SHA Mingyang，ZHANG Sijia，FU Qingcai，et al. Multi-model integrated event extraction for aquatic animal disease prevention and control based on dynamic weight[J]. Journal of Huazhong Agricultural University，2023，42(3)：80-87.

[9] WADDEN D，WENNBERG U，LUAN Yi，et al. Entity，relation，and event extraction with contextualized span representations[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong：[s. n.]，2019：5784-5789.

[10] EBNER S，XIA P，CULKIN R，et al. Multi-sentence argument linking[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]：[s. n.]，2020：8057-8077.

[11] 段佳俊，陆中，王捷. 基于多尺度 CNN 和 Transformer 的航空发动机寿命预测方法[J/OL]. 航空动力学报(2024-11-19)[2024-12-08]. https：//link. cnki. net/doi/10. 13224/j. cnki. jasp. 20240311.
DUAN Jiajun，LU Zhong，WANG Jie. Aero-engine life prediction method based on multi-scale CNN and Transformer[J/OL]. Journal of Aerospace Power(2024-11-19)[2024-12-08]. https：//link. cnki. net/doi/10. 13224/j. cnki. jasp. 20240311.

[12] PRESS O，SMITH N A，LEWIS M. Train short，test long：Attention with linear biases enables input length extrapolation[EB/OL]. (2022-04-22)[2024-12-08]. https：//arxiv. org/abs/2108. 12409.

[13] PENG Tianshuo，LI Zuchao，ZHANG Lefei，et al. FSUIE：A novel fuzzy span mechanism for universal information extraction[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto：[s. n.]，2023：16318-16333.

[14] 凌霞. 妇科圣手杨宗孟临床 56 年经验集[M]. 北京：中国中医药出版社，2016.

[15] 张淑亭，刘桂宇. 张淑亭延嗣医案[M]. 石家庄：河北科学技术出版社，2002.

[16] 中国中医研究院西苑医院. 钱伯煊妇科医案[M]. 北京：人民卫生出版社，2005.

[17] 罗媛媛，杨春明，李波，等. 基于高置信度伪标签数据选择算法的临床事件抽取方法[J]. 太原理工大学学报，2024，55(1)：204-213.
LUO Yuanyuan，YANG Chunming，LI Bo，et al. A clinical event extraction method based on a high-confidence pseudo-label data selection algorithm[J]. Journal of Taiyuan University of Technology，2024，55(1)：204-213.

[18] 杜洁，骆力明，孙众. 基于 ALBERT 预训练模型的事件抽取技术研究[J]. 计算机工程与科学，2023，45(4)：711-717.
DU Jie，LUO Liming，SUN Zhong. Event extraction technology based on ALBERT pre-trained model[J]. Computer Engineering & Science，2023，45(4)：711-717.

[19] 喻雪寒，何琳，徐健. 基于 RoBERTa-CRF 的古文历史事件抽取方法研究[J]. 数据分析与知识发现，2021，5(7)：26-35.
YU Xuehan，HE Lin，XU Jian. Extracting events from ancient books based on RoBERTa-CRF[J]. Data Analysis and Knowledge Discovery，2021，5(7)：26-35.

[20] 张江英，郝矿荣，王直杰，等. 基于 Lattice LSTM-CRF 模型的中文紧急事件抽取[C]//2020 中国自动化大会(CAC2020). 上海：中国自动化学会，2020：770-775.

[21] STRUBELL E，VERGA P，BELANGER D，et al. Fast and accurate entity recognition with iterated dilated convolutions[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen：[s. n.]，2017：2670-2680.