科学技术与教育 2026年第6期
ISSN: 3079-4455

基于融合语义特征的医疗事件抽取模型研究

李晓雪*，刘瑶，贾白茹
陕西国际商贸学院信息工程学院，陕西西安，中国
*通讯作者

【摘要】中文电子病历中的医疗事件提取任务面临专业术语密集、上下文语义捕获不充分等挑战。为缓解上述问题，本文设计了一种融合上下文语义信息与触发词特征的事件提取方案。具体流程包括：利用 LTP 工具完成病历文本的分词处理，并实现对肿瘤电子病历中触发词的自动识别；借助 RoBERTa 预训练模型生成富含上下文语义的词向量表示；进一步，提出基于 BERT-ConditionalLayerNorm 的语义特征融合模型，将全局上下文与局部触发词特征进行联合建模。在真实数据集上的实验结果显示，所提方法有效提高了医疗事件抽取的 F1 指标，证明了该融合策略在缓解语义缺失问题上的有效性。

【关键词】中文电子病历；医疗事件抽取；上下文语义信息；RoBERTa；语义特征融合
【基金项目】陕西省教育厅科研计划项目（编号：24JK0333)；陕西国际商贸学院专项研究项目（编号：SMXY202462）；陕西国际商贸学院校级研究项目(SMXY25042)；陕西国际商贸学院校级研究项目(SMXY25043)

1.引言

作为临床诊疗过程的数字化载体，电子病历（Electronic Medical Record，EMR）在现代医疗体系中扮演着日益重要的角色。它不仅是医护人员记录患者诊疗信息的工具，更是一个动态累积的临床知识库。从数据构成来看，电子病历一方面记录了患者的人口学信息、主诉、现病史、体格检查结果等基础健康信息，另一方面也蕴含着丰富的临床医学知识，包括诊断结论、用药记录、手术操作、检查检验报告等[1-3]。这些多维度、时序化的数据资源，为研究者探究患者健康状态的动态演变规律，以及分析疾病的发生、发展与转归过程，提供了不可或缺的关键数据支撑。

深度学习技术的快速发展推动了事件抽取领域的研究进展。在通用领域，王俊等[4]引入多语义特征融合策略，采用精细化的文档级建模方式提升了事件检测效果；朱敏等[5]则通过双重注意力机制强化了对文本中关键语义单元的捕捉能力。针对医疗场景，余杰等[6]设计了面向中文医疗文本的联合抽取框架，实现了触发词及其相关论元的同步抽取。在生物医学领域，Yu 等[7]构建了基于 LSTM 的端到端框架用于生物医学事件抽取；Fan 等[8]则利用深度学习方法从开放数据中检测和提取不良药物事件。此外，预训练语言模型（如 BERT、RoBERTa、ERNIE）在医疗文本处理中展现出强大能力；Sun 等[9]提出的 OptimalMEE 框架通过微调与后验验证，进一步优化了大语言模型在医疗事件抽取任务上的表现。Gururangan 等[10]的研究表明，在目标领域语料上进行持续预训练能够显著提升下游任务的性能，这为医疗事件抽取的领域自适应提供了重要思路。

现有方法在面向肿瘤等特定疾病的触发词识别环节中，仍较多依赖于人工设计的特征模板，尚未形成完善的自动化语义增强机制。此外，多数模型在处理局部触发词特征与全局上下文表征的融合问题时，倾向于采用向量拼接或逐元素相加等简单策略，难以充分刻画二者之间复杂的深层交互关系[11,12]。基于上述分析，构建一个既能充分利用电子病历上下文语义、又能有效整合触发词特征的事件抽取模型，成为亟待解决的研究问题。

为解决上述挑战，本文构建了一个融合语义特征的医疗事件抽取框架。该框架的核心贡献体现在三个方面：（1）分词与触发词识别：采用 LTP 工具完成病历文本分词，并针对肿瘤病历设计了专门的触发词识别模块。（2）词向量的语义增强：借助 RoBERTa 预训练模型生成词向量，相比传统词嵌入方法，RoBERTa 能更有效地捕捉上下文语义，缓解医学文本中的一词多义问题。

（3）特征融合机制：提出了 BERT-ConditionalLayerNorm 模型，在 BERT 架构基础上引入条件层归一化，将触发词的语义特征作为条件信息融入模型各层，实现全局上下文与局部特征的深度联合建模。

在真实肿瘤电子病历数据集上的实验表明，本模型在医疗事件抽取任务上取得了更优的 F1 值，超越多种基线方法。本研究为电子病历的智能化处理提供了有效工具，也为将领域知识融入预训练模型的研究提供了可参考的思路。

2. SF_BERT 医疗事件抽取模型

本文的主要研究任务是在 CCKS2021 医疗事件抽取评测任务的基础上，重新筛选和标注了 600 条肿瘤电子病历数据，并从中抽取肿瘤的各个属性信息。本文采用了 BIEOS（Begin、Inside、End、Outside、Single）标签系统进行序列标注。

2.1 肿瘤电子病历触发词

2.1.1 触发词识别

基于已标注的肿瘤原发部位信息，筛选出待处理的目标短句；利用 LTP 工具对上述短句执行分词操作，并参照 CCKS2021 官方实体词表，剔除与原发部位共现频次低于 20 的词语[13]；针对语料中可能出现的新词及同义表达，对官方词表进行了人工补充与更新；最终将候选词汇归纳为“癌”“结节”“占位”等 5 个语义类别，形成触发词集合。

表 1 统计了各类触发词的出现频次，这些词汇可作为识别肿瘤病历特征的线索。具体处理流程如下：对病历文本分句后，逐句检测触发词存在情况——包含触发词的句子标记为正例（含有原发部位），否则为负例；最终将筛选出的正例句子送入后续模型训练。图 1 直观展示了上述流程。

2.1.2 触发词分类

确定触发词后，采用决策树算法评估其识别准确度。该算法根据句子中是否包含特定触发词构造决策树。以 500 个数据点为训练集、167 个为测试集，表 2 显示决策树方法取得了较高的识别准确率。

表 1.触发词词频统计

触发词	数量
“癌”	553
“结节”	51
“占位”	39
“影”	21
“团块”	20

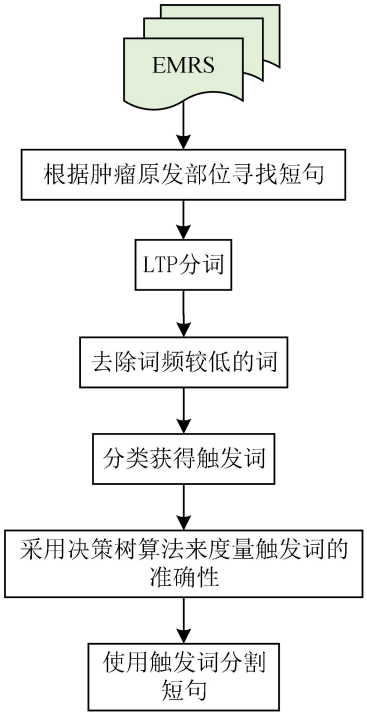

图 1.触发词识别流程

表 2.决策树分类精度

方法	Accuracy
ID3	0.9829
C4.5	0.9829
CART	0.9700

采用 ID3 算法构建决策树：以信息增益最大化为准则选择根节点属性，递归划分子集直至满足停止条件，最终将叶节点标记为分类结果。定义样本集 S 的信息增益为：

$$Gain(\mathrm{S},\mathrm{a}) = \mathrm{H}(\mathrm{S}) - \sum_{v=1}^{V} \frac{|S^v|}{|S|} H(\mathrm{S}^v) \quad (1)$$

$$H(\mathrm{S}) = -\sum_{k=1}^{x} p_k \log_2 p_k \quad (2)$$

2.2 SF_BERT 模型

2.2.1 Roberta 模型

为充分捕捉文本上下文与语义关联，肿瘤事件抽取模型采用 RoBERTa 构建词向量。预训练语言模型在大规模肿瘤病历语料上无监督训练后，可获取更丰富的文本特征与上下文信息，从而提升抽取准确率和泛化能力[14]。图 2 的 BERT 架构通过上述机制增强了模型对上下文的理解，改善了标签预测性能。

将 RoBERTa 应用于肿瘤事件抽取，有助于捕获与肿瘤相关的语义线索及上下文依赖关系，从而提升抽取精度与模型泛化水平。基于该预训练模型，系统能够更精准地从肿瘤病历中解析出各类属性信息，为后续

事件抽取分析奠定良好基础。

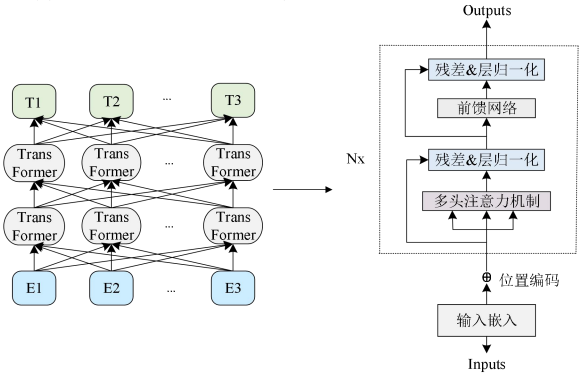

图 2.BERT 模型结构

2.2.2 语义特征与 BERT 模型融合方法

本文设计的 SF_BERT（Semantic-Fused BERT）模型以条件层归一化（Conditional Layer Normalization）为核心机制，该机制依据触发词的相对距离动态调节层归一化操作中的增益系数与偏置项[15]。目标医疗事件属性的抽取由 BERT-ConditionalLayerNorm 架构完成。该模型的输入包含两部分：预处理后的电子病历文本，以及触发词（或关键词“转移”）在文本中的位置编码。通过同时接收这两类信息，模型能够并行获取全局语境特征与局部位置特征，并利用条件层归一化技术实现模型的规范化处理。融合语义特征的医疗事件抽取模型如图 3 所示。

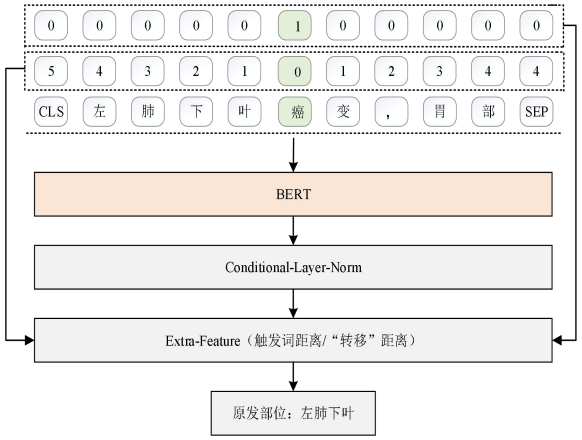

图 3. SF_BERT 医疗事件抽取模型

模型根据任务类型选用不同语义特征：原发部位与病灶大小抽取时，以触发词与目标实体的距离为特征；转移部位抽取时，以“转移”与目标实体的距离为特征。该机制增强了模型对目标实体与其邻接词汇位置关系的捕捉能力。解码器最终将特征映射为事件属性预测结果。

ConditionalLayerNorm 负责将电子病历文本与触发词（或“转移”）的语义信息相融合，并显式标注关键词位置以辅助事件抽取。特征选择上：原发部位和病灶大小采用触发词相对距离；转移部位采用“转移”相对距离。

2.3 医疗事件抽取算法

该算法（见表 3）的处理逻辑如下：从原始文本中筛选出含有触发词或指定关键词的病历短句；对于每个入选短句，计算各词汇与触发词之间的相对位置距离，将其作为语义特征；通过 ConditionalLayerNorm 机制将该语义特征与 RoBERTa 生成的词向量进行融合；接着，利用 RoBERTa 编码器提取融合后的上下文表征，经由线性映射层和 BIEOS 解码器输出医疗事件的相关属性，涵盖原发部位、病灶尺寸以及转移部位等信息。

表 3. SF_SERT Medical Event Extraction Algorithm

Input: text: 原始电子病历文本（字符串） trigger_words: 触发词列表（如“癌”、“结节”等） keyword: 特定语义关键词（原发部位抽取用 trigger_words，转移部位抽取用“转移”）
Output: events: 抽取的医疗事件属性（如原发部位、病灶大小、转移部位）
1 sentences ← split_into_sentences(text)
2 for each sent in sentences:
3 if contains_trigger(sent, trigger_words) or contains_keyword(sent, keyword):
4 target_sentences.append(sent)
5 trigger_pos ← locate_keyword (sent, keyword)
6 for each sent in target_sentences:
7 tokens ← LTP_tokenize(sent)
8 token_ids ← RoBERTa_tokenizer(tokens)
9 rel_distances ← compute_relative_distance (trigger_pos, each_token_position)
10 semantic_features ← embed(rel_distances)
11 for each token_id, sem_feat in zip(token_ids, semantic_features):
12 word_embed ← RoBERTa_embed(token_id)
13 conditioned_embed ← ConditionalLayerNorm (word_embed, sem_feat)
14 context_representation ← RoBERTa_encoder (conditioned_embed)
15 logits ← Linear(context_representation)
16 predicted_labels ← argmax(logits, dim=-1)
17 events ← decode_BIEOS(predicted_labels, tokens)
18 return events

3.实验分析

3.1 模型参数及评价指标

模型参数指标如表 4 所示，模型性能评估采用准确率（P）、召回率（R）及微观平均 F1 值三项指标，具体计算方式见式（3）-（5）。其公式如下：

表 4.模型参数设置

参数	数值
Batch Size	32
Epoch	100
Learning _ rate	2e-4
dropout	0.2
Bert Embedding size	768
Sequence Length	160
Optimizer	Adam

$$P=\frac{TP}{TP+FP} \quad (3)$$

$$R=\frac{TP}{TP+FN} \quad (4)$$

$$F1=\frac{2\cdot p\cdot r}{p+r} \quad (5)$$

对比实验以中文 RoBERTa-wwm-ext-large 为基准。泛化能力评估采用 10 折交叉验证方案：每折数据迭代训练 10 轮，总计 100 次实验，依据式（6）计算 F1 均值；再将各折内 10 轮结果的标准差取平均值，即为式（7）所示的平均标准差。

$$F1_{mean}=\frac{\sum_{i=0}^{9}\sum_{j=0}^{9}F1_{i,j}}{100} \quad (6)$$

$$SD_{mean}=\frac{\sum_{i=0}^{9}SD_i}{10} \quad (7)$$

3.2 语义特征对模型的影响

表 5 显示：LTP 与 BERT-ConditionalLayerNorm 的融合方案在肿瘤病历数据上效果良好，加入语义特征后 F1 值明显上升；10 折交叉验证则缓解了类别不均衡问题，增强了模型泛化能力。

表 5.融合语义特征的医疗事件抽取模型结果

模型	F1 值
BERT- ConditionaLayerNorm	72.23%
特征+BERT- ConditionaLayerNorm	78.01%
10 折交叉验证	78.89%

3.3 模型对比分析

基于 RoBERTa-wwm 的实验表明，语义特征的逐步融入有效提升了模型 F1 值。以 BiLSTM-CRF 、 BiLSTM-Attention-CRF 和 Transformer-BiLSTM-CRF 为基线，本模型在融入语义特征后 F1 值分别高 7.35%、7.65% 和 5.98%（见表 6）。上述结果证实，本方法在医疗事件抽取任务上具有明显的性能优势，为电子病历的智能化处理提供了一种有效且可行的技术路径。

表 6.医疗事件抽取模型对比

模型	F1 值
BiLSTM-CRF	70.66%
BiLSTM-Attention-CRF	70.36%
Transformer-BiLSTM-CRF	72.03%
特征+BERT-ConditionaLayerNorm	78.01%

3.4 误差分析

肿瘤医疗事件抽取采用管道式架构，将流程拆分为多个可独立优化的步骤。然而，该架构存在误差累积问题——前序触发词识别的偏差会向后传递[16]。本研究中，触发词识别处于流程前端，其误差将扩散至特征融合与事件抽取环节，最终损害整体抽取效果。

4.结语

针对医疗电子病历中专业术语密度高、现有事件抽取模型对上下文语义建模能力不足等问题，本文提出了一种融合语义特征的预训练语言模型方案。该方案的核心创新在于：将触发词的语义特征作为条件信息融入 BERT 架构，实现全局上下文与局部特征的联合建模。训练过程中引入十折交叉验证机制，有效提升了模型的泛化能力。

在真实肿瘤电子病历数据集上的实验结果表明，融入语义特征后模型的 F1 值获得显著提升，验证了该方法在缓解上下文语义缺失问题上的可行性与实用价值。与 BiLSTM-CRF、BiLSTM-Attention-CRF、Transformer-BiLSTM-CRF 等基线模型相比，本模型在准确率和召回率方面均展现出明显优势。需要指出的是，本研究目前存在一定局限性：实验数据主要采集自单一医疗机构，且集中于肿瘤这一特定疾病类型，模型在更广泛医疗场景下的适用性有待进一步验证。此外，模型的计算复杂度较高，在实际临床部署中可能需要考虑轻量化改造。

未来工作将围绕以下三个方向展开：一是开展多源、多病种的跨机构数据验证，检验模型的泛化能力；二是研究模型压缩与知识蒸馏技术，降低计算开销，便于临床部署；三是探索文本与医学影像等多模态信息的融合路径，进一步提升事件抽取的准确性。通过这些改进，有望将本模型推广至更广泛的医疗智能化应用场景。

参考文献

[1] Deimazar G, Sheikhtaheri A .Machine learning models to detect and predict patient safety events using electronic health records: A systematic review [J]. International journal of medical informatics, 2023(Dec.):180.
[2] 梁文桐，朱艳辉，詹飞，等.基于伪标签置信选择的半监督医疗事件抽取[J].微电子学与计算机，2022，39（01）：71-79.
[3] 马连韬，张超贺，焦贤锋，等.Dr. Deep：基于医疗特征上下文学习的患者健康状态可解释评估[J].计算机研究与发展，2021, 58(12)：2645.
[4] 王俊，史存会，张瑾，等.融合上下文信息的篇章级事件时序关系抽取方法[J].计算机研究与发展，2021，58（11）：2475-2484.
[5] 朱敏，毛莺池，程永，等.基于双重注意力机制的事件抽取方法[J].软件学报，2022：1-15.
[6] 余杰，纪斌，刘磊，等.面向中文医疗事件的联合抽取方法[J].计算机科学，2021，48（11）：287-293.
[7] Yu X, Rong W, Liu J, Zhou D, Ouyang Y, and Xiong Z. LSTM-Based End-to-End Framework for Biomedical Event Extraction [J]. IEEE Transactions on Computational Biology and Bioinformatics, 2021, 22(1):194-207.
[8] Fan B, Fan W, Smith C. Adverse drug event detection and extraction from open data: A deep learning approach [J]. Information Processing & Management, 2020, 57(1): 102131.Medicine.Springer, Cham, 2024.
[9] Sun Y, Wang S, Li Y, et al. Ernie 2.0: A continual pre-training framework for language understanding [A]. Proceedings of the AAAI conference on artificial intelligence [C]. 2020, 34(05): 8968-8975.
[10]Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks [J]. ACL, 2020:8342-8360.
[11]Fócil-Arias C, Sidorov G, Gelbukh A, et al. Extracting medical events from clinical records using conditional random fields and parameter tuning for hidden Markov models [J]. Journal of Intelligent & Fuzzy Systems, 2018, 34(5):2935-2947.
[12]Jiang B, Zhu S, Wu J, et al. A Joint Learning Framework for Document-Level Event Extraction [J]. IEEE Transactions on Neural Networks and Learning Systems, PP[2026-04-24].
[13]Botnar K, Nguen T J, Farnsworth G M, et al. EHRchitect: An open-source software tool for medical event sequences data extraction from Electronic Health Records [J]. Journal of clinical and translational science, 2025, 9(1):e79.
[14]Sun Y, Wu D, Chen Z, et al. OptimalMEE: Optimizing Large Language Models for Medical Event Extraction through Fine-Tuning and Post-hoc Verification [C]//International Conference on Artificial Intelligence in Medicine. Springer, Cham, 2024.
[15]李晓雪.面向中文电子病历的医疗事件抽取[D].西北师范大学，2023.
[16]An B, Zhang H, Ma L, et al. End-to-end Chinese clinical event extraction based on large language model [J]. Scientific Reports [2026-04-23].