传染病流调报告智能解析：基于深度学习的 TBEE 模型及其应用

胡 健[1,2] 焦增涛[3] 蔡康宁[2,3] 梁万年[1,4]

（[1] 清华大学万科公共卫生与健康学院 北京 100084 [2] 清华大学生物医学工程学院 北京 100084
[3] 医渡云（北京）技术有限公司 北京 100083 [4] 清华大学健康中国研究院 北京 100084）

〔摘要〕 **目的/意义** 基于深度学习算法构建传染病流调报告智能解析模型，以提升公共卫生应急响应能力。**方法/过程** 结合疾控业务需求，系统分析流调报告内容要素，提出基于 BERT 语义编码和 Bi－LSTM 时序建模的改进模型 TBEE，以实现流调报告的高效结构化处理。**结果/结论** 该模型在事件抽取任务中的实体级 *F*1 分数超过 80%，显著优于 TMT－NN、Bi－LSTM＋CRF 及 Llama3 等对比模型。能够在不依赖高性能计算条件的情况下，为快速获取传染病传播关键线索、提升数据分析能力和应急处置效率提供有力支持。

〔关键词〕 流调报告；大语言模型；深度学习；公共卫生应急响应
〔中图分类号〕 R－058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2025.12.011

Intelligent Analysis of Infectious Disease Epidemiological Investigation Reports: the TBEE Model Based on Deep Learning and Its Application

HU Jian[1,2], *JIAO Zengtao*[3], *CAI Kangning*[2,3], *LIANG Wannian*[1,4]

[1] *Vanke School of Public Health, Tsinghua University, Beijing 100084, China*; [2] *School of Biomedical Engineering, Tsinghua University, Beijing 100084, China*; [3] *Beijing Yidu Cloud Technology Co. Ltd., Beijing 100083, China*; [4] *Institute of Healthy China, Tsinghua University, Beijing 100084, China*

〔**Abstract**〕 **Purpose/Significance** To construct an intelligent analysis model for infectious disease epidemiological investigation reports based on deep learning algorithms, so as to enhance the capacity for public health emergency response. **Method/Process** Combined with the business requirements of disease control, the content elements of the epidemiological investigation reports are analyzed, and an improved model TBEE based on BERT semantic coding and Bi－LSTM time series modeling is proposed, achieving efficient and structured processing of epidemiological investigation reports. **Result/Conclusion** Experimental results demonstrate that this model achieves an entity－level *F*1 score exceeding 80%, significantly outperforming compared models including TMT－NN, Bi－LSTM＋CRF, and Llama3. It can provide strong support for the rapid acquisition of key clues about the spread of infectious diseases, the improvement of data analysis capabilities and the efficiency of emergency response without relying on high－performance computing conditions.

〔**Keywords**〕 epidemiological investigation report; large language model; deep learning; public health emergency response

〔修回日期〕 2025－10－09

〔作者简介〕 胡健，博士研究生，发表论文 20 篇；通信作者：梁万年，教授。

〔基金项目〕 科技创新 2030——“新一代人工智能”重大项目（项目编号：2021ZD0114100）；北京重大呼吸道传染病研究中心课题（项目编号：BJRID2025－014）。

1 引言

流调报告作为追踪传染病传播路径、识别密切接触者与风险场所的关键依据，其处理效率和准确性对公共卫生应急响应具有重要意义。传统流调报告多以非结构化文本形式记录患者的时空轨迹与行为信息，依赖人工解读，存在效率低、易遗漏、难复用等问题。当传染病传播快速蔓延时，结构化处理任务成为提升研判质量和处置效率的核心环节。

随着自然语言处理（natural language processing，NLP）技术的发展，从非结构化文本中自动提取信息成为可能，尤其是命名实体识别（name entity recognition，NER）、关系抽取和事件抽取等任务，已取得显著突破。在流调报告中，关键信息通常包括患者身份、活动轨迹、接触对象及时间地点等。事件抽取任务可将文本内容转换为结构化事件元组，支持传播链路分析与风险区域识别。然而，传统模型多基于通用语料训练，对领域特定表达与事件逻辑建模不足。

双向编码器表示模型（bidirectional encoder representations from transformers，BERT）[1-2] 作为预训练语言模型代表，能够捕捉深层语言信息，广泛应用于命名实体识别、关系抽取等任务。双向长短期记忆网络（bidirectional long short－term memory，Bi－LSTM）及其衍生的模型（如 dbRNN[3]、JMEE[4] 和 RCEE[5]）是一种改进的循环神经网络结构，能同时从前向和后向两个方向学习文本序列信息，从而更全面地理解词语在上下文中的含义，适用于描述患者行程轨迹等时间敏感任务。随着公共卫生信息化水平提升，大量研究致力于从流调报告、电子病历和临床文本中提取结构化信息以辅助公共卫生决策。Wang J 等[6] 提出 TMT－NN 方法，基于 BERT 结合实体识别与规则推理机制，构建结构化事件表示，支持传播路径可视化。在事件抽取方面，Raza S 等[7] 基于 Bi－LSTM 和条件随机场（conditional random field，CRF）模型进行传染病流调事件抽取，并取得良好效果。Truhn D 等[8] 借助 GPT－4 实现医学报告结构化，证明大语言模型在文本结构化方面的能力，但其部署成本较高，实际落地受限。这些方法在结构化能力方面提供了理论支持与技术基础，但在事件粒度、时序建模及资源适配方面仍存在不足，见表 1。因此，本研究提出基于时间的神经网络信息抽取（time－aware bi－directional event extractor，TBEE）模型，融合 BERT 编码与 Bi－LSTM 结构，结合时间段注意力机制和全局调控评分策略，强化模型对流调报告中事件顺序与关联的建模能力，实现多事件元组的准确识别。基于真实中文流调报告数据集开展实验，结果显示该方法在 *F*1 指标上优于 TMT－NN、Bi－LSTM＋CRF 等传统模型，且在无图形处理器（graphic process unit，GPU）环境下仍展现出良好的处理效率，具备在一线应用场景中推广的潜力。

表 1 相关方法对比

方法	主要应用/场景	工作原理要点	识别对象/输出	成效概述	局限/风险	资源需求（相对）
TMT－NN	流调报告文本结构化	规则/词典＋模型；结合知识图谱	时间、地点、人物等基本实体	利用已有词典数据，结构化能力强	依赖知识图谱维护；对跨句依赖较弱	中（需知识图谱服务）
Bi－LSTM＋CRF	通用序列标注	上下文序列编码＋标签转移约束	人物、地点、时间等基本实体	轻量、易部署	难以显式建模跨段时序关系	低（CPU 可用）
Llama3 8B	通用大模型问答/抽取	大语言模型通过少样例学习完成新抽取任务	根据自然语言指令灵活抽取时间、地点、症状等对象	跨语料/任务抽取泛化能力较好	资源消耗大，稳定性依赖提示设计	高（至少具备单卡高显存 GPU）
TBEE（本研究）	流调事件抽取与时序建模	BERT＋时间段位置编码＋自注意力＋Bi－LSTM＋全局分数	时间＋事件＋动作链条	实体级 *F*1 ≥ 80%；跨日/多事件链条解析稳定	长文本极端情况召回率降低	低、中（单卡消费级 GPU 或 CPU 加速）

2 资料与方法

2.1 研究方法

流调报告通常以患者基本信息为起始段，以时间为关键词，表述患者在不同时间段的行程及其密切接触者。流调报告中的患者信息书写格式通常较规范，可以通过简单的正则表达式处理，甚至可通过填表等形式在信息录入初期就完成结构化处理。因此本研究重点为事件抽取。融合 BERT 编码和 Bi－LSTM 结构，提出以时间为关键信息的抽取方法。该方法通过以下 7 步实现对流调报告中事件实体的精准提取，见图 1。

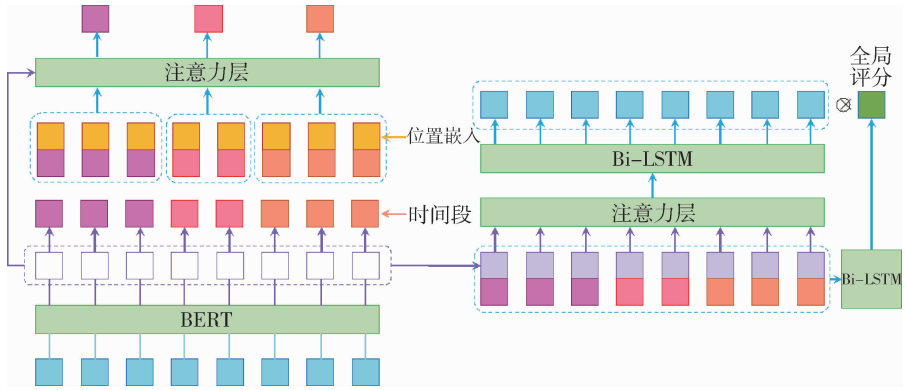

图 1 TBEE 模型结构

第 1 步：文本预处理与 BERT 编码。将原始流调报告文本输入 BERT 中，以 BERT 为文本编码器，得到流调报告编码向量。文本输入记为 $Text = \{w_1, w_1, \cdots, w_n\}$，经 BERT 编码后得到上下文向量表示 T，其中 T_{ti} 为不同时间段落文本。

$$T = BERT(Text) \tag{1}$$

$$T = [T_{t1}, T_{t2}, \cdots, T_{tn}] \tag{2}$$

第 2 步：时间段位置嵌入及段落聚合。对不同时间段落文本 T_{ti} 分别加入位置编码 P_t，再进行池化操作，得到查询向量 q。

$$q_{ti} = pooling(T_{ti} + P_t), i = 1, 2, \cdots, n \tag{3}$$

$$Q = [q_{t1}, q_{t2}, \cdots, q_{tn}] \tag{4}$$

第 3 步：注意力机制特征选择。通过注意力机制提取不同时间段关注的信息，注意力层中的键（key，K）、查询（query，Q）、值（value，V）用于计算注意力分数并实现全局评分调控，其中 W_k 为键投影矩阵，W_v 为值投影矩阵，d 代表 K 和 Q 向量的维度，a 代表注意力机制计算后得到的输出向量。

$$K = W_k T \tag{5}$$

$$V = W_v T \tag{6}$$

$$a_{ti} = \frac{q_{ti} K^T}{\sqrt{d}} V, i = 1, 2, \cdots, n \tag{7}$$

第 4 步：联合上下文与原始表示。将注意力机制抽取的特征 A 与相应编码文本 T 拼接起来，作为信息抽取的输入。其中 A 代表 a_{ti} 特征矩阵。

$$A = [a_{t1}, a_{t2}, \cdots, a_{tn}] \tag{8}$$

$$input = concat(T, A) \tag{9}$$

第 5 步：自注意力操作。将处理后的输入信息经过自注意力操作，使特征向量根据当前任务需要，动态地从全局信息中捕捉与自身最相关的上下文特征。

$$Q = W_q input \tag{10}$$

$$K = W_k input \tag{11}$$

$$V = W_v input \tag{12}$$

$$hidden_{attention} = \frac{Q K^T}{\sqrt{d}} V \tag{13}$$

第 6 步：Bi－LSTM 时序建模。通过 Bi－LSTM 识别文本中患者参与的事件。

$$pos = Bi - LSTM(hidden_{attention}) \tag{14}$$

第 7 步：全局评分。为了使模型具有全局观念，对输入特征求得全局分数 globalscore，微调模型的输出概率分布，从而提高最终识别性能。

$$globalscore = Bi - LSTM(input) \tag{15}$$

$$pos = globalscore \times pos \tag{16}$$

这些步骤可分为两部分，第 1 部分为时间段感知位置编码，其主要目的是捕捉流调报告文本中跨时段的事件演化关系，以时间为关键信息，为每一词引入其所属时间段的嵌入向量，将其与 BERT 输出向量相加，使模型具备对事件发生时间的建模能力，有效识别“同日多事”或“跨日传播”场景下的关联实体。第 2 部分为 BERT + Bi – LSTM 联合编码结构，先使用 BERT 编码器提取上下文丰富的词语表示，再通过 Bi – LSTM 捕捉词序依赖信息，从而增强对时间顺序敏感事件描述的理解能力，适用于流调报告中典型的行程轨迹类表述。在模型训练过程中，由于采用了预训练语言模型 BERT 作为基础，初始化已较充分，因此不采用交叉验证方式进行超参数确定。模型以划分好的固定验证集作为超参数验证依据，并使用 6 个轮数设定对模型进行充分训练。

2.2 数据集

采用 Wang J 等[6] 共享的数据集，其包含 2 264 条由中国疾病预防控制中心与主流媒体公开发布的流调报告中文文本，涵盖发病、确诊、就医、入院、出院等 8 类事件类型，以及时间、症状、地点、交通工具等 7 种事件元组要素。为构建模型训练所需结构化输入，将原始报告按时间线索（如“1 月 25 日”“当天”等）划分为多个段落，然后参考数据集中官方定义的 BIO 标注体系（将文本中每个词标记为实体开头 B、实体内部 I 或非实体 O），对段落中涉及的时间、位置、交通工具、症状等实体进行序列标注。按 8∶1∶1 的比例划分训练集、验证集与测试集，其中训练集 1 811 条、测试集 227 条、验证集 226 条。数据集实体类型，见表 2。

表 2 数据集实体类型

实体类型及数量	具体实体
8 类 type	Event：患者产生的各类活动对应的动词。Onset：疾病发病事件对应的动词。HospitalVisit：患者就医事件对应的动词。DiagnosisConfirmed：患者确诊事件对应的动词。Inpatient：患者入院事件对应的动词。Discharge：患者出院事件对应的动词。Death：患者死亡事件对应的动词。Observed：患者被作为疑似病例观察事件对应的动词
7 类 tuple	Date：时间或者日期。Symptom：症状。LabTest：实验室检测。ImagingExamination：影像检查。Location：省市级别的位置。Spot：地点，如宾馆、酒店、家等。Vehicle：交通工具，如火车、汽车等

2.3 实验方法

为验证 TBEE 模型的效果，选取目前主流事件抽取算法（TMT – NN、Bi – LSTM + CRF 和 Llama 3 8b[9]）作为对比。其中 TMT – NN 是专门针对疾控流调报告事件的抽取算法；Llama 3 8b 是包含 80 亿参数的大语言模型，可捕捉广泛的一般知识，在信息抽取和自然语言理解任务中效果较好[8]。针对本研究特定临床分类任务，对 Llama 3 8b 采用低秩适配器（low – rank adaptation，LoRa）方法，以实现高效微调。该方法有效减轻了微调工作量，并允许快速适应新任务。对于微调过程，使用预先训练好的权重和 LoRa 适配器，其中秩 $r = 16$，缩放系数 $\alpha = 16$。

采用自然语言处理中的 3 项常用指标评估模型表现。精确率指模型判定为正例的实体中实际为正例的比例；召回率指所有真实实体中被模型成功找出的比例；*F*1 分数是精确率与召回率的调和平均数，用于反映模型的综合性能。所有指标均基于实体级别进行计算，通过对比数据集的标注结果，分别评估模型在各类实体上的抽取准确性与覆盖能力。

3 实验结果

3.1 模型对比结果

TBEE 与 3 种对比模型在测试集上的结果，见表 3。不同模型对具体流调报告事件抽取效率的对比结果，见表 4。TBEE 模型在流调报告数据集上显示出较好性能，对于绝大部分关键信息能够做到准确识别。与 TMT – NN 模型和 Bi – LSTM + CRF 模型相比，TBEE 模型在精确率和 *F*1 上均取得最优效

果，在抽取的时间颗粒度、交通/地点准确度、行为链条完整度、跨句一致性以及资源效率方面具有明显优势。Bi－LSTM＋CRF 模型在召回率上略优，但是精确率明显低于 TBEE 模型。与 Llama3 8b 微调后模型相比，TBEE 在准确率和召回率上均领先，基于 BERT 架构的模型在信息抽取任务上比 GPT 架构的模型效果更好，而且 Llama3 需要专门的 GPU 服务器，计算成本较高。

表 3 不同模型信息抽取结果对比（%）

模型方法	精确率	召回率	*F*1
TMT－NN	77.6	80.0	78.8
Bi－LSTM＋CRF	75.1	84.2	79.6
Llama3 8B	81.2	82.4	81.8
TBEE	84.1	83.0	83.6

表 4 不同模型资源与效率对比

模型方法	最小硬件	处理时长/报告	输出稳定性	典型差错
TMT－NN	CPU＋知识图谱服务	中	取决于规则覆盖	跨句顺序弱
Bi－LSTM＋CRF	CPU	低（快）	序列边界稳定	跨段信息缺失
Llama3 8B	单卡高显存 GPU	高（慢）	受提示词影响	过度生成/不一致
TBEE	单卡消费级 GPU 或 CPU	低－中（较快）	跨时序一致性强	极长文本召回有待提升

3.2 消融实验结果

通过消融实验对比按时间信息分段提取的特征和原始文本特征拼接与否的结果，见表 5，以验证模型结构的合理性。结果表明，利用时间特征对 BERT 的输出进行处理有利于流调报告文本信息抽取，精确率、召回率、*F*1 值均得到显著提高。针对部分与时间具有直接关联性的实体类别（症状、地点和交通工具），单独进行消融实验，见表 6。结果表明，加入时间特征可大幅提升这些类型实体的召回率，从而使整体 *F*1 得分明显提升。

表 5 消融实验结果（%）

模型方法	精确率	召回率	*F*1
TBEE	84.1	83.0	83.6
消融实验－无时间特征处理	82.9	69.6	75.7

表 6 部分实体消融实验结果（%）

模型方法	实体类别	精确率	召回率	*F*1
TBEE	Location	83.2	79.0	81.0
	Symptom	91.1	82.7	86.7
	Vehicle	81.7	79.2	80.3
消融实验－无时间特征处理	Location	81.1	71.2	75.5
	Symptom	90.7	74.3	81.7
	Vehicle	82.5	70.7	76.2

4 讨论

4.1 模型结构的针对性优势

TBEE 模型的优异表现源于其结构设计与任务特征的契合。时间位置编码显著增强了模型区分“同日多事件”的能力，避免了上午、下午、晚间事件的错配；全局评分机制保证了跨句、跨段的一致性，使“人物－地点－时间”的组合关系更稳定；Bi－LSTM 在序列建模中捕捉了事件的先后顺序，并与自注意力机制互补，从而在长文本和复杂链条中保持较高识别准确率。消融实验结果进一步表明，去除时间编码或全局评分均导致性能下降，这说明各模块设计均有效。

4.2 时序与不确定性信息识别优势

流调文本常包含“可能”“疑似”等模糊表述，传统模型对此往往直接忽略。TBEE 在建模过程中引入时间段位置编码与置信度机制，使模型既能区分不同时间片段的事件，又能对不确定事件进行合理标注。在实验中，TBEE 在涉及模糊表述的样本上召回率明显优于其他模型，显示出其在实际流调任务中处理复杂语句的优势。

4.3 资源适配与部署优势

与参数规模庞大的 Llama3 相比，TBEE 模型具有轻量化特征，对硬件环境要求低。在 4 核 Intel i5 CPU 上即可实现每分钟约 38 份流调报告的处理速度，而人工处理等量报告平均需 29 分钟。这表明 TBEE 不仅提升了效率，也具备低资源场景下的实用性。要指出的是，GPU 在本研究中仅用于提升批量推理速度，并非文本处理所必需的条件，因此 TBEE 在资源受限的应急环境中更具部署优势。

5 结语

流调报告的快速结构化处理是公共卫生应急管理中极为关键的环节。患者基本信息往往通过表格或标准化形式获得，较易实现结构化，而复杂的行程事件信息难以直接抽取。为此，本研究提出 TBEE 模型，在事件抽取任务中实体级 *F*1 分数超过 80%，显著优于 TMT - NN、Bi - LSTM + CRF 和 Llama3 等对比方法。该模型能够有效实现实际流调报告的结构化处理；其轻量化设计进一步保障了在公共卫生应急响应体系中的可推广性，从而为提升研判效率、节约人力与物资成本提供了有力的技术支撑。然而，本研究仍存在一定局限。一是采用流水线结构可能导致时间实体识别阶段的误差向后续事件链条传递；二是模型在处理超长文本和识别稀有事件时的召回率仍有提升空间。未来研究可探索端到端的统一建模方法，并引入层次化文本结构理解，以减少误差传播，进一步增强模型在复杂现实场景下的鲁棒性与泛化能力。

作者贡献：胡健负责研究设计、文献调研、实验实施、论文撰写；焦增涛负责研究设计、提供指导、论文审核；蔡康宁负责研究设计、文献调研、实验实施；梁万年负责研究设计、提供指导、论文审核与修订。
利益声明：所有作者均声明不存在利益冲突。

参考文献

1 DEVLIN J, CHANG M W, LEE K, et al. BERT: pre - training of deep bidirectional transformers for language understanding [C]. Minneapolis: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

2 VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. [2025 - 08 - 23]. https://arxiv.org/abs/1706.03762.

3 SHA L, QIAN F, CHANG B, et al. Jointly extracting event triggers and arguments by dependency - bridge RNN and tensor - based argument interaction [C]. New Orleans: The Thirty - Second AAAI Conference on Artificial Intelligence, 2018.

4 LIU X, LUO Z, HUANG H. Jointly multiple events extraction via attention - based graph information aggregation [C]. Brussels: The 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.

5 LIU J, CHEN Y, LIU K, et al. Event extraction as machine reading comprehension [C]. Online: The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

6 WANG J, WANG K, LI J, et al. Accelerating epidemiological investigation analysis by using NLP and knowledge reasoning: a case study on COVID - 19 [C]. Washington D. C.: The AMIA Annual Symposium, 2020.

7 RAZA S, SCHWARTZ B. Entity and relation extraction from clinical case reports of COVID - 19: a natural language processing approach [J]. BMC medical informatics and decision making, 2023, 23 (1): 20.

8 TRUHN D, LOEFFLER C M, MÜLLER - FRANZES G, et al. Extracting structured information from unstructured histopathology reports using GPT - 4 [J]. Journal of pathology, 2024, 262 (3): 310 - 319.

9 Meta. Introducing Meta Llama 3: the most capable openly available LLM to date [EB/OL]. [2025 - 10 - 05]. https://ai.meta.com/blog/meta - llama - 3/.