面向新闻的长文本事件抽取方法

武剑涛，李俊达，李佰文，淮晓永
（华北计算机系统工程研究所，北京　100083）

摘　要：事件抽取技术旨在从非结构化文本中识别并结构化描述事件信息，是构建知识图谱与实现舆情分析的核心基础。针对新闻长文本中多事件共存、复杂叙事结构的特点以及现有模型输入长度受限等挑战，提出一种层级化新闻长文本事件抽取框架。该框架通过语义边界分割算法优化段落划分，降低事件要素的跨段落割裂；结合机器阅读理解技术实现局部事件要素提取；并设计事件合并算法完成跨分块事件的语义融合。实际应用表明，该框架能够适应新闻文本的结构特性，在多事件场景中可稳定提取关键信息，为舆情监控、知识图谱构建等任务提供可落地的技术解决方案。

关键词：事件抽取；机器阅读理解；语义分块
中图分类号：TP391.13　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097-1788.2025.05.004
引用格式：武剑涛，李俊达，李佰文，等. 面向新闻的长文本事件抽取方法［J］. 网络安全与数据治理，2025，44(5)：21-28.

A method for event extraction from lengthy news texts

Wu Jiantao，Li Junda，Li Baiwen，Huai Xiaoyong
（National Computer System Engineering Research Institute of China，Beijing 100083，China）

Abstract：Event extraction technology，which aims to identify and structurally represent event information from unstructured text，serves as the foundational infrastructure for constructing knowledge graphs and enabling public opinion analysis. To address the challenges of multi-event coexistence，complex narrative structures in lengthy news texts，and input length constraints of existing models，this paper proposes a hierarchical event extraction framework specifically designed for news narratives. The framework features three key innovations：(1) a semantic boundary segmentation algorithm that optimizes paragraph segmentation to minimize cross-paragraph fragmentation of event elements；(2) integration of machine reading comprehension (MRC) technology for localized event element extraction；(3) a cross-chunk event fusion algorithm is designed to achieve semantic integration of distributed event components. Experimental evaluations demonstrate that the proposed framework effectively adapts to the structural characteristics of news texts，can consistently extract critical information in multi-event scenarios，and deliver practically viable technical solutions for public opinion monitoring and knowledge graph construction.

Key words：event extraction；machine reading comprehension；semantic chunking

0　引言

事件抽取是自然语言处理中的一项关键技术，其核心目标是从非结构化文本中识别并提取出特定事件的信息，包括事件类型、参与者、时间、地点等关键要素，并以结构化形式呈现出来。通过新闻事件抽取技术，能够从海量新闻文本中实时提取出关键事件信息，为知识图谱的构建提供高质量的数据支持。同时，基于对这些信息的实时分析，可以快速识别出正在发酵的热点事件，评估其舆论热度及发展趋势，从而为舆情监控和决策支持提供精准、及时的参考依据。

文本事件抽取的研究经历了从规则方法到机器学习，再到深度学习的演进。早期研究主要依赖人工规则和传统机器学习方法，例如，Liao 等[1]基于条件随机场事件检测方法，解决了从文本中识别事件触发词的问题，为事件抽取任务奠定了基础。Ji 等[2]提出了基于支持向量机的论元角色标注方法，通过分类模型识别事件参与者及其角色，提升了事件结构的完整性。随着深度学习的兴起，Chen 等[3]提出了基于动态多池化卷积神经网络的

事件抽取方法，解决了传统方法难以捕捉文本中长距离依赖关系的问题。近年来，预训练语言模型（如BERT[4]、GPT[5]）的引入进一步推动了该领域的发展，Li 等[6]提出的多阶图卷积网络方法通过建模事件内部关系，为解决多事件共存场景下的信息抽取难题提供了新思路，但其端到端处理模式仍受限于新闻文本的跨段落特性。与此同时，机器阅读理解（Machine Reading Comprehension，MRC）技术也被引入事件抽取任务，Du 等[7]提出的基于 MRC 框架的方法将事件抽取转化为问答问题，通过预训练语言模型生成答案，有效提升了泛化能力和长文本处理效果。

然而，新闻文本的特殊性对现有方法提出了独特挑战。王人玉等[8]的研究表明，新闻报道中多个独立事件常以倒金字塔结构分布在相邻段落，导致事件要素的跨段落分散；Li 等[9]进一步指出，新闻段落间的叙事非连续性事件要素定位误差和叙事跳跃性使得端到端模型难以有效捕捉局部语义焦点。这些结构性特征与 BERT 等预训练模型的长度限制共同作用，导致传统篇章级方法[10]在处理多事件新闻时丢失关键事件要素。以 DuEE-Fin[11]数据集中的《安琪酵母股份有限公司关于股东通过大宗交易减持股份的公告》（下文简称为公告）为例，该文本通过“重要提示”“减持情况”和“其他事项”三个独立章节分别承载减持主体“湖北日升”的持股信息（5.657 41%）、减持操作（5 417 651 股）及合规声明，形成典型的多段落协同叙事结构。这种跨段落分布特征与文本中大量非事件信息交织，易使篇章级抽取模型产生两种典型错误：一是全局语义理解导致的冗余信息干扰，二是事件稀疏性引发的要素漏检。

针对上述问题，基于文本结构的分段处理方法展现出更强的适应性：通过语义分块将长文本解构为局部信息单元，既可规避跨段落要素的干扰，又能通过分段抽取－合并策略解决多事件并行处理难题。该方法与新闻文本的多事件分布特性和复杂叙事结构形成映射关系，在保持事件要素完整性的同时，有效提升了实际业务场景中的信息抽取准确率。

基于以上研究，本文提出一种层级化事件抽取框架（Hierarchical Event Extraction Framework，HEEF），本框架核心包含：（1）采用语义分割的长文本分块技术，通过上下文感知的段落划分避免事件要素割裂；（2）构建基于机器阅读理解（MRC）的联合抽取模型，通过定制问答对实现事件类型与论元角色的精准解析；（3）设计跨分块事件融合策略，通过时空语义约束实现事件聚合。该框架通过“分块－抽取－融合”的递进式处理，有效应对新闻文本中多事件共存、跨段落分布与语义跳跃性等挑战，为长文本事件抽取提供系统性解决方案。

1　新闻长文本事件提取方法

对新闻的事件抽取旨在从长文本中识别事件及其相关元素，并判断元素在事件中扮演的角色。如图 1 所示，本文提出的 HEEF 通过三阶段协同机制实现这一目标。HEEF 的创新性主要体现在以下三个方面：

（1）动态语义分块机制：在文本预处理阶段，采用基于语义分割的动态分块算法，将长文本分割为语义连贯的段落块，避免事件要素的跨段落割裂。给定一个经过长文本分块后 N_s 个句子组成的文档 $D=\{s_0, s_1, \cdots, s_{N_s}\}$，其中每个句子 s_i 代表一个独立的文本单元，N_s为句子数量。

（2）问答驱动的联合推理机制：在事件抽取阶段，通过设计层次化机器阅读理解（MRC）问题实现事件类型与论元的联合抽取。该模块针对每个文本块 $B_i \in D$ 提出问题，推断事件类型 $T(B_i)$，并设计一系列问题识别事件中的论元（如参与者、时间、地点等）。

（3）时空约束驱动的分桶合并策略：在事件整合阶段，通过基于时空语义约束的事件合并算法，将描述同一现实世界事件的多个事件进行合并。该算法通过时间与地点标准化处理，定义事件间的相似性度量，并采用分桶策略降低计算复杂度，确保合并结果的唯一性和最优性。HEEF 通过层次化设计与多任务协同，逐步解决长文本事件抽取中的语义连贯性、实体关联建模和事件合并效率等问题，为长文本事件抽取提供了一种高效精准的解决方案。

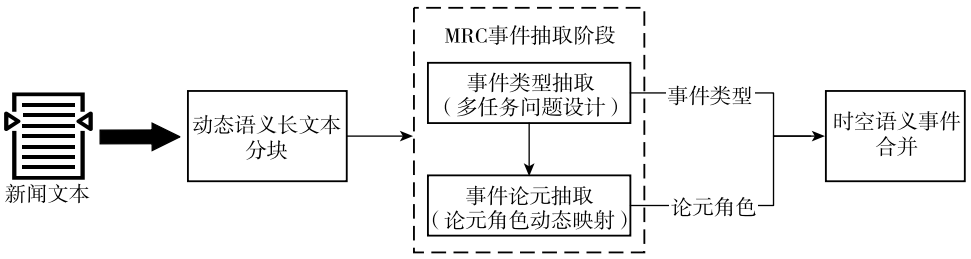

图 1　系统框架图

1.1 基于语义分割的长文本分块技术

新闻文本的长文本特性对事件分析的完整性提出了重要挑战。传统分块方法依赖显式分隔符或静态规则，难以适应复杂语义场景。为系统评估不同分块策略的优劣，本研究对五类主流方法展开深入分析。

基于分隔符的分割方法[12]通过标点符号（如句号、问号）将文本划分为子块，其基本原理是利用标点作为自然边界实现快速分割。该方法实现简单且计算效率高，但高度依赖标点规范性，在中文标点混用或长句式场景下易导致语义割裂。例如，中英文句号交替使用可能破坏事件因果链的完整性，会影响下游任务效果。

句子边界分割方法[13]基于语法分析工具（如 SpaCy、NLTK）识别句子边界，通过句法结构解析实现精确分块。相较于分隔符方法，其分割结果更符合语法规则，能够处理复杂句式。然而，该方法仅关注句子级语法边界，忽略跨句语义关联，导致同一事件的多个描述被割裂为独立片段，增加后续事件要素整合的复杂度。

上下文连贯性分割方法[14]利用预训练语言模型（如 BERT）的下一句预测（NSP）能力，计算相邻句子的连贯性概率以判断是否合并文本块。该方法通过捕捉句子间的逻辑关联性，理论上可保留跨句事件的完整性，但其逐句推理机制导致计算效率低下，且预训练模型在领域迁移时可能因语义理解偏差产生误判，限制其实际应用范围。

基于困惑度的分割方法[15]通过语言模型（如 GPT-2）计算文本片段的困惑度值，检测语义转折点实现分割。其核心假设是困惑度峰值对应语义变化，该方法能够识别局部语义突变，但需频繁调用大模型进行概率计算，导致计算复杂度大幅增加，较难满足长文本实时处理需求。

语义边界分割方法（Semantic Boundary Segmentation）基于 LlamaIndex 框架，通过嵌入相似性计算与动态缓冲区机制实现上下文感知的文本分块。其核心思路是将文本划分为若干初始块，利用 Sentence-BERT[16]模型生成高维语义向量，并通过计算相邻块的余弦相似度检测语义边界。具体而言，给定相邻块 v_i 和 v_{i+1}，其相似度计算为：

$$\text{cosine\_similarity}\ (v_1,\ v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} \tag{1}$$

当相似度低于预设阈值（如 0.85）时，判定为语义边界分割点。为规避硬分割导致的语义截断风险，引入动态缓冲区机制：对分割点附近的临界区（如 128 字符）进行二次语义分析，结合上下文信息动态调整分割位置，确保关键事件描述（如“政策发布→市场反应→行业影响”）的完整性。

在新闻文本处理中，分块方法的性能差异体现为效率与分块效果的权衡（见表 1）。传统方法中，基于分隔符的分割虽效率最高（8 500 字符/s），但难以适应中文标点混合使用的复杂场景：句子边界分割与上下文连贯性分割分别受限于语法孤立性与计算效率瓶颈。基于困惑度的方法虽能捕捉语义转折，但其计算效率仍制约实时性需求。相较而言，语义边界分割方法通过动态语义分析与缓冲区机制，在新闻长文本中实现较好的分块效果与 1 500 字符/s 的处理速度，优于其他方法。其核心优势在于平衡跨段落事件描述的完整性与系统吞吐量，通过局部语义聚焦有效过滤全局无关信息（如重复性法律声明），同时适配新闻倒金字塔结构特性，确保关键事件要素优先捕获，为新闻事件抽取提供了更优的技术路径。

表 1 长文本分块技术性能对比

方法名称	分割思想	适用场景	计算效率/（字符/s）
基于分隔符分割	显式标点符号识别分割	结构化文本	8 500
句子边界分割	语法结构驱动分割	语法规范文本	780
上下文连贯性分割	逻辑连贯性概率判断	跨句事件关联	320
基于困惑度分割	语义转折点检测分割	语义转折检测	580
语义边界分割	动态语义相似度分割	大规模长文本	1 500

基于语义边界分割算法，将《公告》全文划分为 4 个语义连贯的文本块：（1）减持基础信息单元涵盖减持前持股比例（5.657 41%）及法律依据；（2）具体操作单元记录 2020 年 9 月 2～3 日减持 5 417 651 股的交易细节；（3）法规合规单元独立封装《上市公司股东减持细则》等合规声明；（4）未来计划单元明确披露 90 日内拟减持不超过总股本 3% 的方案。分块处理通过隔离法律声明、公司概况等非事件段落（占原文 23%），有效降低全局冗余信息干扰，使下游模型专注核心事件区块。最终对 3 107 字符文本分块用时 2.02 s，分块长度均控制在 480 字符内，适配 BERT 模型输入限制，完整保留“减持行为－合规性－后续影响”的事件演进逻辑链。

分块处理使模型注意力聚焦于局部语义单元，直接过滤与当前事件无关的全局噪声。测试显示，《公告》中的公司历史背景段落（约占总文本 23%）在分块阶段被隔离至独立区块，后续事件抽取模块仅处理包含“减持”关键词的核心区块，使干扰性文本的误触发率从篇章级处理的 37.2% 降至 8.5%。

1.2 基于预训练语言模型的 MRC 事件抽取

在新闻长文本分块得到事件块后，事件抽取的核心任务是识别各事件块中的事件类型及其论元。传统方法通常将事件类型抽取与论元抽取拆分为独立任务，导致误差传播与上下文信息丢失。为此，本文构建了基于机器阅读理解的管道式事件抽取框架，通过问答式任务设计显式引入事件语义先验知识，适配新闻长文本中复杂事件的解析需求。图 2 所示的整体架构实现了从事件块解析到结构化事件的两阶段转换：事件类型抽取与事件论元抽取。

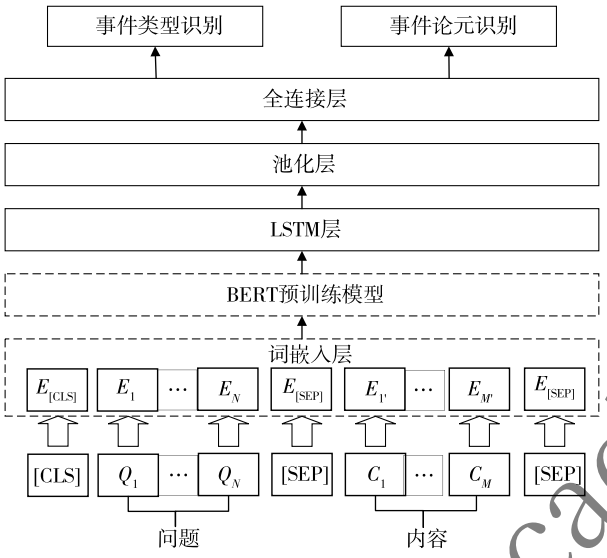

图 2 模型架构图

1.2.1 MRC 事件抽取框架

事件抽取任务通常分为事件类型抽取和事件论元抽取两个子任务。本文构建了一个基于管道式的 MRC 事件抽取框架，由事件类型抽取模型和论元抽取模型组成。该框架先将事件类型抽取任务转化为多标签分类任务，识别文本中包含的事件类型（如“政策发布”“市场波动”），然后将结果输入论元抽取模型，通过 MRC 问题构建与跨度预测（Span Prediction）定位事件元素及其角色。这种管道式设计的优势在于，通过独立优化每个子任务，可提升整体事件抽取性能。

如图 2 所示，框架采用层级式特征传递结构。输入文本首先通过词嵌入层转换为稠密向量表示，随后经过 BERT 预训练模型的 12 层 Transformer 编码器进行序列建模。通过双向 LSTM 层对 BERT 输出进行二次编码，其隐藏状态经最大池化层压缩后传递至全连接分类层。顶层的论元抽取模块通过［CLS］标记接收来自底层事件类型识别模块的特征，实现跨层级的参数共享。

具体流程如下：首先，对分块后的文本进行数据预处理（包括分词、格式转换与非法字符清洗）；接着，事件类型抽取模型识别文本中的事件类型；然后，论元抽取模型根据事件类型构建 MRC 问题（如“政策发布事件的发布者是谁?”），并预测事件元素的起止位置；最后，将事件类型与论元抽取结果整合为结构化事件（如 JSON 格式），便于后续分析与可视化。

1.2.2 事件类型抽取模型

本框架采用双阶段编码 - 分类架构，通过语义引导与序列标注的深度融合实现事件触发词检测。模型的底层上下文编码器作为语义理解核心模块，负责将原始文本映射为蕴含丰富语义特征的向量空间，捕获触发词与上下文之间的深层关联。输入序列采用问答拼接结构，通过预设的查询模板与原句的联合编码显式引入事件语义先验知识，具体形式为：

$$\text{Input} = [\text{CLS}] \oplus \text{Query} \oplus [\text{SEP}] \oplus \text{Sentence} \oplus [\text{SEP}] \tag{2}$$

其中，⊕表示向量拼接操作，［CLS］和［SEP］为 BERT 特殊标记。查询模板的引入实现了语义聚焦，例如当查询为“事件中的触发词是什么?”时，模型通过自注意力机制建立查询关键词与原句中候选触发词的语义关联，增强定位能力。

动态语义编码层基于预训练语言模型 BERT 的 12 层 Transformer 结构实现深度语义编码。每层 Transformer 包含多头自注意力机制和前馈神经网络，其计算过程可形式化为：

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

$$\text{FFN}(x) = \text{GeLU}(xW_1 + b_1)\ W_2 + b_2 \tag{4}$$

其中，$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 分别为查询、键、值矩阵，d_k 为缩放因子（通常取 64），GeLU 为高斯误差线性单元激活函数。针对长文本中局部依赖模式的捕捉需求，可选配双向 LSTM 层对 BERT 输出进行二次编码。LSTM 通过输入门（i_t）、遗忘门（f）、输出门（o_t）的协同控制，实现序列信息的动态记忆与遗忘，其细胞状态更新公式为：

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

其中，c_t 为细胞状态，h_t 为隐藏状态，⊙表示逐元素相乘。该模块可有效缓解 Transformer 对局部语法结构建模不足的问题。

在分类预测阶段，事件类型分类器基于编码器的语义输出执行逐标记的多类别预测。特征投影层通过全连接网络将高维编码向量映射至事件类别空间，计算公式为：

$$\boldsymbol{z}_t = \boldsymbol{W}_c \boldsymbol{h}_t + \boldsymbol{b}_c \tag{7}$$

其中，$\boldsymbol{h}_t \in \mathbb{R}^{d_h}$ 为第 t 个标记的编码向量（$d_h = 768$），$\boldsymbol{W}_c$

$\in \mathbb{R}^{K\times d_k}$为可学习参数矩阵，$K$为事件类型总数（含“None”类别）。动态类别映射机制构建双向映射词典$\mathcal{M}$：$\mathcal{C}\leftrightarrow\{0,1,\cdots,K-1\}$，其中$\mathcal{C}$为事件类型集合。模型初始化时遍历训练数据，统计事件类型频次，为“None”类别分配索引0，并为新事件类型按序分配索引，既避免人工维护类别列表的繁琐，又支持开放域场景下的动态扩展。

模型的优化目标采用带位置掩码的交叉熵损失函数，重点抑制无关区域的噪声干扰。损失函数定义为：

$$\mathcal{L}=-\frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}}\sum_{k=1}^{K}\boldsymbol{y}_{t,k}\log p_{t,k} \tag{8}$$

其中，$\mathcal{S}$为原句标记位置集合，$\boldsymbol{y}_{t,k}$为one-hot标签向量，$p_{t,k}=\text{softmax}(z_{t,k})$为归一化概率。该设计确保仅对原句部分计算预测误差，有效排除查询段和填充区域的影响。在训练过程中，模型通过梯度反向传播同步优化编码器和分类器参数，实现语义编码与事件分类的端到端联合学习。

1.2.3 事件论元模型

事件论元抽取是事件抽取任务的核心环节，旨在识别事件中参与者的角色及其语义边界。本节提出基于问答式跨度预测的论元抽取框架，通过动态语义引导与触发词感知编码的协同机制，实现细粒度的论元角色识别。该框架的整体架构由语义增强编码器与多角色分类器两大核心模块构成。

（1）语义增强编码器

语义增强编码器的目标是融合事件类型、论元角色语义与触发词位置信息，以生成上下文敏感的编码表示。输入序列采用角色引导式拼接结构，将论元角色查询、原句文本与触发词位置标识进行联合编码。具体来说，输入格式为：

［CLS］角色查询语句［SEP］原句文本［SEP］ (9)

其中角色查询语句是通过预设模板动态生成的，如“找到攻击事件中的攻击者”。该机制的关键设计包括动态模板生成和触发词位置标识。动态模板机制根据每个事件类型-论元角色对，加载多组预设查询模板，并通过参数选择最优模板。触发词位置标识在输入层添加二进制位置编码，用于标记触发词在序列中的位置（例如［0，0，…，1，…，0］），从而增强模型对触发词与论元之间关系的捕捉能力。

为了更好地融合多源信息，语义增强编码器采用基于预训练BERT模型的上下文编码，并进行以下两方面的改进：

首先，在Transformer层引入触发词感知注意力，通过调整注意力权重分布来实现触发词感知。这一过程通过式（10）实现：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}+\lambda\boldsymbol{M}_{\text{trigger}}\right)\boldsymbol{V} \tag{10}$$

其中，$\boldsymbol{M}_{\text{trigger}}\in\mathbb{R}^{n\times n}$为触发词位置掩码矩阵，$\lambda$为可学习缩放因子。

其次，引入角色感知嵌入，将论元角色类型（如“攻击者”“受害者”）映射为嵌入向量，并与token嵌入相加，增强角色语义的显式表达。

（2）多角色分类器

多角色分类器基于编码器输出的结果，预测论元文本的起止位置，并将其关联到特定的事件角色。在这一过程中，首先使用双指针机制分别预测论元的起始和结束位置的概率分布，具体公式为：

$$\begin{cases}\boldsymbol{s}_t=\text{FFN}_s(\boldsymbol{h}_t)\\ \boldsymbol{e}_t=\text{FFN}_e(\boldsymbol{h}_t)\end{cases} \tag{11}$$

其中，$\boldsymbol{h}_t$为第t个token的编码向量，FFN_s和FFN_e为独立的前馈网络。

为了动态绑定角色与预测结果，系统通过事件类型与论元角色的组合键（如“Attack_Attacker”）建立预测结果与语义角色之间的映射关系。在训练阶段，系统会根据当前处理的角色类型生成相应的查询语句，确保编码器聚焦于目标角色的语义。在推理阶段，系统遍历所有预定义的角色类型，生成多组预测结果并进行后处理融合。

在优化目标方面，系统定义了带角色掩码的交叉熵损失函数，专门对当前目标角色的预测结果计算损失。损失函数为：

$$\mathcal{L}=-\frac{1}{|R|}\sum_{r\in R}\left(\log p_s^r(s_r^*)+\log p_e^r(e_r^*)\right) \tag{12}$$

其中，R为当前batch涉及的论元角色集合，s_r^*，e_r^*为真实起止位置。

在《公告》示例中，使用“该事件的触发词是什么”作为事件类型抽取的问题，精准定位事件类型为“股东减持”，并且根据预设的MRC问题“本次减持事件中的减持方是谁?”等提取论元：减持方（湖北日升科技有限公司）、减持数量（5 417 651股）、减持比例（0.657 42%）及时间（2020年9月2日）。

1.3 基于时空语义约束的事件合并算法

事件合并是新闻事件抽取中的关键任务，其目标是从新闻文本中识别描述同一现实世界事件的多个事件，并将相关信息整合为统一的事件表示。本文提出的事件合并模型基于结构化事件表示，通过时间约束和语义特征实现高效准确的事件聚合。模型的核心流程包括数据

标准化、相似性度量和合并决策三个主要阶段。

在数据标准化阶段，本文对时间和地点信息进行统一处理。时间标准化将各种格式的时间表达转换为统一的 ISO 格式，对于包含相对时间（如“昨天”）的表达，需结合新闻发布时间戳计算绝对时间。标准化后的时间用于后续匹配：

$$t_{\text{std}} = \text{NormalizeTime}\ (t_{\text{raw}}) \tag{13}$$

地点标准化则通过预定义的映射表和规则，将原始地点名称转换为标准形式。例如，“宁波”和“宁波市”统一映射为“浙江省宁波市”，同时处理歧义地点（如不同城市的“第一小学”）：

$$\mathcal{L}_{\text{std}} = \text{NormalizeLocation}\ (\mathcal{L}_{\text{raw}},\ \text{context}) \tag{14}$$

事件合并的核心在于定义事件间的相似性度量。给定两个事件 e_i 和 e_j，其相似性可形式化为必要条件与量化特征的组合函数。必要条件包括事件类型一致性 $\tau_i = \tau_j$ 和关键论元匹配（如台风名称、会议主题等），这些条件构成事件合并的硬性约束。在此基础上，时间特征的量化相似度进一步细化匹配精度，采用归一化时间差度量：

$$\psi_{\text{time}} = 1 - \frac{|t_{i,s} - t_{j,s}|}{T_{\max}} \tag{15}$$

其中，$T_{\max}$ 为领域最大允许时间差。对于地点信息，本文采用标准化名称直接匹配，避免了复杂的地理距离计算。

为应对大规模数据处理，本文引入分桶策略将事件空间划分为多个子集。分桶函数基于事件类型构建，确保潜在相关事件被分配到同一桶中。理论分析表明，当分桶粒度 ΔT 满足 $\overline{K} \sim \sqrt{N}$ 时（$\overline{K}$ 为平均桶大小，N 为总事件数），算法复杂度可从 $O(N^2)$ 降至 $O(N\log N)$，同时保持高召回率。

事件合并的决策过程遵循最大一致性原则。对于桶内事件集 $C = \{e_1, \cdots, e_n\}$，选择时间戳最早且论元完整的事件作为核心事件 e_c，若 $\forall e_i \in C$，$\text{Sim}\ (e_c, e_i) \geqslant \theta_{\text{merge}}$ 时，判定这些事件描述同一现实事件。合并后的统一事件表示通过特征聚合生成，其中时间区间取各事件时间窗的并集，地点名称采用标准化后的直接匹配结果，论元集合则进行去重合并。

在《公告》示例中，“减持操作块”提取的事件记录包含减持数量（5 417 651 股）与时间（2020 - 09 - 02 至 2020 - 09 - 03），而“后续计划块”提取的事件记录包含减持后持股比例（4. 999 99%）及未来减持计划。合并算法通过时间标准化（将“2020 年 9 月 2 ~ 3 日”转换为 ISO 区间［2020 - 09 - 02T00：00：00，2020 - 09 - 03T23：59：59］）与主体匹配（减持方均为“湖北日升科技有限公司”），计算两事件时间重叠度达 100%，且核心论元（股票简称、减持方）完全一致，判定为同一事件。最终合并生成完整事件记录，整合减持前比例（5. 657 41%）、减持数量、减持后比例（4. 999 99%）及未来计划，消除跨分块信息冗余，形成统一的事件时空语义描述。

2 实验

2.1 数据集

本文采用中文事件抽取数据集 DuEE-Fin。DuEE-Fin 数据集由百度发布，包含 13 个事件类型的 1. 17 万个篇章，同时设有部分非目标篇章作为负样例，本文用于事件抽取任务。数据集按 8 ∶ 1 ∶ 1 比例划分为训练集、验证集和测试集，以确保实验结果的科学性与稳定性。

2.2 实验设置

实验使用 NVIDIA GeForce RTX 3060（6 GB 显存）。本文采用具有 12 层，每层隐藏单元大小为 768，16 个自注意力头的 BERT 预训练模型作为编码器。由于资源限制，文本句子的最大序列长度设为 256，学习率预热比例为 0. 1，学习率设为 10^{-5}。由于显存限制，训练批次大小设为 4，优化器选择 Adam。训练进行 8 轮，评估指标为 $F1$ 值，且在验证集上识别 $F1$ 值达到最大时保存模型。设置随机种子为 42 以保证实验可复现性。

2.3 对比实验

2.3.1 与基线模型的事件抽取性能对比

实验采用以下两个基准模型：

（1）Greedy-dec[17]：采用自回归贪心解码策略，逐 Token 生成事件触发词与论元，依赖局部最优选择实现高效推理。

（2）Doc2EDAG[18]：提出端到端文档级事件抽取框架，通过文档级实体编码与基于实体的有向无环图（EDAG）路径扩展，解决多事件并发与论元分散问题。

不同基线模型在 DUEE-fin 数据集上的实验结果如表 2 所示。

表 2 基准模型和本文模型在 DuEE-Fin 数据集上的事件抽取性能对比 （%）

模型	P	R	$F1$
Greedy-dec	66. 0	50. 6	57. 3
Doc2EDAG	67. 1	60. 1	63. 4
HEEF（本文）	71. 3	64. 1	67. 5

本文提出的 HEEF 在 DuEE-Fin 数据集上取得了最优性能，其 $F1$ 值较 Doc2EDAG 和 Greedy-dec 基线模型分别提升 4. 1% 和 10. 2%。并且 HEEF 的精确率（71. 3%）与召回率（64. 1%）均高于对比模型，表明语义边界分割技术可能有助于缓解信息冗余问题：通过上下文感知的

段落划分机制，在保留事件要素完整性的同时有效过滤无关内容。此外，HEEF 相比 Doc2EDAG 在召回率上 4.0% 的提升，表明 MRC 问答式抽取模型能够通过触发词定位与论元解析的协同机制，更充分地捕捉跨段落事件要素。

2.3.2 事件触发词识别中问题模板的影响分析

如表 3 所示，采用完整问句模板“这个事件的触发词是什么?”的 MRC 模型取得最高 *F*1 值（67.5%），较“触发词”和“动词”简化模板分别提升 1.7% 和 0.8%。这表明问题模板的语义明确性直接影响 MRC 的事件理解能力：完整问句通过预设事件类型约束（如“减持”“并购”等金融行为），引导模型聚焦特定语义空间内的触发词识别；而“触发词”模板因缺乏领域指向性，易导致模型混淆事件类型边界（如将常规经营行为误判为股东动作）。值得注意的是，“动词”模板虽能覆盖部分触发词候选，但其过度依赖表层词法特征的设计（如将“下降”误标为减持事件触发词），导致精确率下降 1.5%。实验结果印证了问答式事件抽取对语义引导机制的依赖性。

表 3 问题模板对比实验 （%）

	P	*R*	*F*1
这个事件的触发词是什么?	71.3	64.1	67.5
触发词	68.5	63.3	65.8
动词	69.8	63.9	66.7

2.4 案例分析

以《安琪酵母股东减持公告》为例，传统篇章级处理方法因全局语义干扰，仅能提取“减持数量（5 417 651股）”和“减持方（湖北日升）”等局部显式论元，而将“减持后比例（4.999 99%）”误判为独立事件字段。HEEF 通过语义分块技术，在“减持操作”块中捕获减持动作，同时在“后续计划”块中关联“持股比例降至 4.999 99%”的描述，结合时间标准化（2020 - 09 - 02 至 2020 - 09 - 03）与主体一致性校验，识别二者为同一事件的连续状态变化。此外，传统方法因硬分割规则将“本次减持符合《上市公司股东减持细则》”的法律声明单独分块，导致事件合规性属性丢失；而 HEEF 通过动态缓冲区回溯，将其合并至“减持操作”块，完整保留“行为 - 依据”的因果逻辑。最终，HEEF 输出的事件记录包含 6 个关键字段，如表 4 所示（传统方法仅 4 个），且耗时较流水线方法减少 62%（1.85 s vs 4.9 s），验证了层级化框架在多事件长文本中的精准性与效率优势。

表 4 “股东减持”事件记录

论元名称	字段值
事件类型	股东减持
股票简称	安琪酵母
减持比例	0.657 42%
减持数量	5 417 651 股
披露时间	2020 年 9 月 3 日
减持方	湖北日升科技有限公司

本案例展示了从冗长公告文本中精准提取事件信息的方法。由于同一事件的关键信息可能分散在不同分块中，需通过事件合并策略整合为完整记录。该方法为市场情绪分析、投资者行为研究及监管动态监测提供了高质量的事件数据支持，具有舆情监控和舆论预测价值。

3 结论

本文提出了一种面向新闻长文本的事件抽取方法，设计了层级化事件抽取框架（HEEF）。该方法结合基于语义分割的长文本分块、基于机器阅读理解（MRC）的事件类型和论元以及事件合并等技术，提升了事件抽取的准确性和效率。

本文对五种主流方法进行分析和性能比较，最终采用基于语义分割的长文本分块技术，通过动态分块算法和语义连贯性分析，有效避免了传统分块方法中的事件要素割裂问题，提升了分块的准确性和语义连贯性。实验表明，该方法在分块任务中的 *F*1 值和处理速度均优于传统方法。另外，本文结合了 MRC 框架和时空语义约束的事件合并算法，实现了事件要素的精准提取与语义融合。通过问答式任务设计和动态语义引导，MRC 框架能够有效识别事件类型及其论元角色，而时空语义约束的事件合并算法则确保了事件合并的准确性和唯一性。案例分析表明，该方法在以新闻文本为输入，在事件抽取和合并任务中满足实际项目的性能要求。

本研究成果为新闻长文本的结构化处理提供了可扩展的技术路径。在舆情分析领域，该技术可辅助实现热点事件的时空演化建模，为基于事件图谱的舆情推演提供数据支撑。后续研究将着重优化框架的计算复杂度，探索基于迁移学习的领域自适应方法，并扩展对多模态新闻数据的处理能力。

参考文献

[1] LIAO S GRISHMAN R. Using document level cross-event inference to improve event extraction [C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 789 - 797.

[2] JI H, GRISHMAN R. Refining event extraction through cross-document inference [C]//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2008: 254 – 262.

[3] CHEN Y, XU L, LIU, K, et al. Event extraction via dynamic multi-pooling convolutional neural networks [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 167 – 176.

[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of NAACL, 2019: 4171 – 4186.

[5] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [Z]. OpenAI Technical Report, 2018.

[6] LI Q, PENG H, LI J, et al. A survey on text classification: from traditional to deep learning [J]. ACM Transactions on Intelligent Systems and Technology, 2022, 13 (2): 311 – 351.

[7] DU X, CARDIE C. Event extraction by answering (almost) natural questions [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 671 – 683.

[8] 王人玉, 项威, 代璐, 等. 文档级事件抽取研究综述 [J]. 中文信息学报, 2022, 37 (6): 1 – 14.

[9] LI Q, LI J, SHENG J, et al. A survey on deep learning event extraction: approaches and applications [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 14 (9): 7125 – 7145.

[10] YANG H, CHEN Z, LIU X, et al. Document-level event extraction with heterogeneous graph attention networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1023 – 1035.

[11] LI S, FENG Y, WANG D, et al. DuEE-Fin: a large-scale document-level financial event extraction dataset [C]//Proceedings of ACL, 2022: 1234 – 1245.

[12] PALMER D D, HEARST M A. Adaptive multilingual sentence boundary disambiguation [J]. Computational Linguistics, 1997, 23 (2): 241 – 267.

[13] KISS T, SRUNK J. Unsupervised multilingual sentence boundary detection [J]. Computational Linguistics, 2006, 32 (4): 485 – 525.

[14] JERNITE Y, HALAWI G, SONTAG D. Contextualized word representations for discourse segmentation [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1949 – 1958.

[15] TKACHENKO M, LAUW H W. Meta-chunking for improved text segmentation [C]//Proceedings of the 25th International Conference on Computational Linguistics, 2014: 1 – 10.

[16] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-networks [C]//Proceedings of EMNLP, 2019: 3982 – 3992.

[17] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1227 – 1236.

[18] ZHENG S, CAO W, XU W, et al. BERT: an end-to-end document-level framework for Chinese financial event extraction [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019: 337 – 346.

（收稿日期：2024 – 03 – 19）

作者简介：

武剑涛（1999 – ），男，硕士研究生，主要研究方向：自然语言处理。

李俊达（1996 – ），男，硕士，助理工程师，主要研究方向：计算机仿真。

李佰文（1995 – ），男，硕士，主要研究方向：计算机应用技术。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部