基于网络资源的震后应急处置灾情信息可视化平台研究与实现

丁宸炀[1] 黄 猛[1] 任玺睿[1] 邵文博[1]
DING Chenyang HUANG Meng REN Xirui SHAO Wenbo

摘 要 地震应急工作中及时获取各类地震应急处置信息并直观展示，对震后抗震救援工作十分关键。随着信息技术的快速发展，震后网络资源中的各类灾害信息会随着时间变化呈现几何式增长且动态变化，但由于以上处置信息具有文本较长、语义内容繁杂散乱等特点，传统事件信息抽取方法在复杂长距离文本中的多类型事件抽取中容易出现嵌套实体和上下文语义复杂不明确的问题，从而影响到整体事件论元抽取的准确性。针对以上问题，采用基于 Python 的 Selenium 自动化技术 +APScheduler 定时模块进行数据多时段获取自动化获取，以近 10 年的历史灾情信息构建灾情信息语料库，并构建基于机器阅读理解框架，结合 BiGRU 网络的多轮问答式地震应急处置灾情信息抽取模型，对各类地震应急处置事件进行论元抽取。为了方便展示，基于 WEBGL 技术 +VUE+PHP 技术开发了地震应急处置灾情信息平台，在 2023 年多次地震的应用中，所提出的平台模型计算快速准确，为震后应急救援工作提供信息辅助。

关键词 地震应急处置灾情信息；机器阅读理解；事件抽取；信息可视化

doi：10.3969/j.issn.1672-9528.2024.02.037

0 引言

破坏性地震发生后，网络资源中出现的相关地震应急信息数量会随着时间的推移呈现幂次级增长，数据量高达十几万条。以 2022 年 9 月 5 日四川泸定 6.8 级地震为例，震后 12 h 后信息暴增到上百万条，将近增长上千倍，而对震后应急救援工作十分关键的人员伤亡、救援出队、救援物资、道路损坏、房屋损坏等多类地震应急处置信息就隐藏在以上海量信息中。同时，在地震应急处置灾情信息中，可能会出现一些包含多类地震应急处置灾情事件的长文本信息，情况如表 1 所示，比如救援物资和救援出队这两类地震应急处置事件往往容易出现在同一条应急处置信息中，并含着大量的“数字 + 计量单位”片段，此类片段往往结构非常相似，但所对应的实体类型和角色意义并不相同。如果文本距离较长语义较复杂，模型会提取诸如 { 出队人数：268 辆 } 等抽取实体与实体角色不符的错误结果。与此同时，在抽取事件论元时，容易出现实体嵌套识别问题，如“四川省消防救援总队”这一类组织单位实体中包含“四川省”这一类地点实体。

目前主流的事件主要分为三种，分别为基于模式匹配的事件抽取方法、基于机器学习的事件抽取方法和基于深度学习的事件抽取方法。基于模式匹配的事件抽取方法使用人工构造规则进行抽取，后期维护成本较高，而且模式的可移植性差，缺乏语义泛化能力；基于机器学习的事件抽取方法将事件抽取分为多个子任务，分别进行触发词的检测和论元的识别，而对复杂的长距离文本在上游任务中容易出现误差累积的问题，而基于深度学习的事件抽取方法主要是依靠命名实体识别产生的实体信息等。这些方法在很大程度上取决于实体信息进行参数获取。而地震应急处置灾情信息具有信息内容较长且杂糅的特点，传统事件抽取方法在复杂长距离文本中（如表 1）的多类型事件抽取中容易出现嵌套实体和上下文语义复杂不明确的问题，从而影响到整体事件论元抽取的准确性。近年来，一种基于机器阅读理解模式（machine reading comprehension，MRC）在事件抽取上的良好表现为以上问题提供了新的解决思路，利用 MRC 将事件抽取任务转化为问答模型。这种方法利用了深度学习的优势，也减少了对上游任务的依赖，同时解决了地震应急处置灾情长文本信息中的多类事件抽取中论元实体嵌套和语义不清晰的问题。因此，针对以上问题，采用基于 Python 的 Selenium 自动化技术 +APScheduler 定时模块进行数据多时段获取自动化获取 [1]，采用 MRC 机器阅读理解框架对地震应急处置文本信息进行事件抽取，根据对应事件抽取模板采用多轮问答模式引入触发词先验知识编码结合双向 BiGRU 网络。学习合成的地震应急处置文本向量的整体特征，增强上下文语义信息，并对各类地震应急处置灾情事件论元抽取进行起始位置预测，从而实现多类型地震应急处置灾情事件抽取及信息可视化，为震后应急救援工作提供信息辅助。

1. 防灾科技学院 河北三河 065201

表 1 网络资源地震应急处置灾情信息获取样例表

序号	部分网络资源地震应急处置灾情信息获取样例	来源
样例 1	9 月 5 日 12 时 52 分，甘孜州泸定县境内发生 6.8 级地震。地震发生后，由 1086 名消防救援人员、268 辆消防车、17 头搜救犬、34 艘舟艇组成的消防救援力量在四川省消防救援总队统一指挥调度下，争分夺秒奔赴灾区，在泸定和石棉的深山峡谷、高空激流中跋山涉水，开展一次次生命营救。	中国新闻网
样例 2	截至 9 月 6 日 17 时，四川全省高速公路收费站应急通道已保障 1200 余辆应急救援车辆免费快速通行，派出抢险救援队伍 19 支，共计 734 人次，以及挖掘机、装载机等救援机具设备 169 台班，已成功抢通多处阻断点。地震发生后，四川交通调集储备应急客运车辆 560 辆、应急货运车辆 466 辆，已运送人员 800 余人次、物资近 200 吨；组织调派船舶 6 艘，运送伤员及救援队伍 120 余人；随时准备根据震区需求增调其他市（州）应急运力。	中国运输交通部官网新闻

1 系统介绍

震后为了对海量网络资源中的地震应急处置灾情信息进行多时段自动化动态获取，本系统爬取近 10 年的历史灾情信息，构建灾害信息语料库，并结合自然语言处理模型技术进行人员伤亡、救援出队、道路损毁等多类地震应急处置灾情信息的精准抽取分析及时空化展示，搭建具有动态性、实时性及直观性的震后应急处置灾情信息可视化平台。本项目采用了基于 python 的 Selenium 自动化技术 +APScheduler 模块 +MySQL 数据库的方法进行地震应急处置灾情信息的结构化存储 [2]，应用 Element 技术结合 Arcgis 三维地图实现地震应急处置灾情信息数据时空可视化。同时基于机器阅读理解框架的地震应急处置信息事件抽取模型，根据对应事件抽取模板采用多轮问答模式引入触发词先验知识编码结合双向 BiGRU 网络。学习合成的地震应急处置文本向量的整体特征，增强上下文语义信息，提高了事件论元抽取精度，最终实现地震应急处置信息三维可视化。一共分为三个模块，地震应急信息可视化模块、地震应急处置灾情新闻事件管理模块和震例数据管理模块。

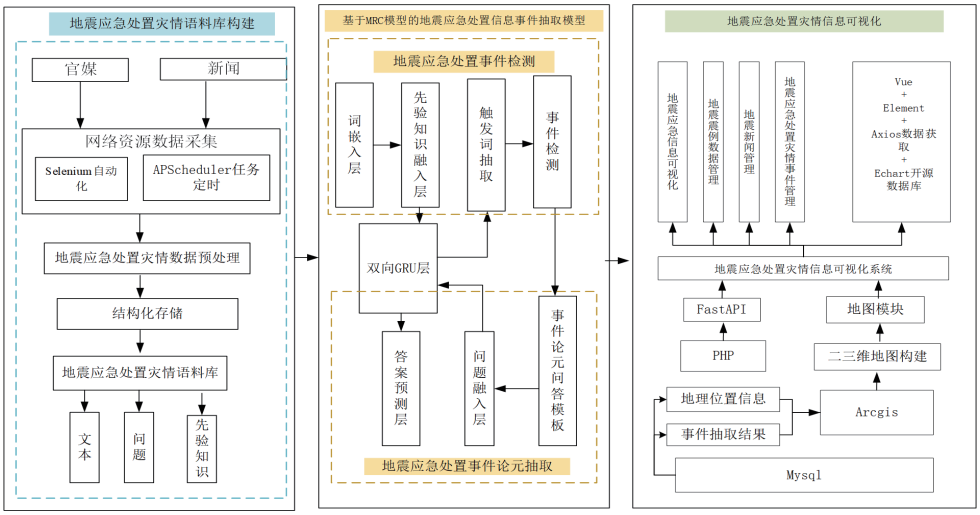

图 1 系统技术路线图

2 系统整体架构概述

2.1 系统技术路线

本系统旨在将自然语言处理和地震应急结合起来，采用微服务架构方式，以 VUE 为前端开发框架，以基于 Python 的 FsatAPI，PHP 作为后端开发，结合 ajax 请求及 CORS 资源开宇共享进行前后端交互及应用开发。首先使用基于 python 的 Selenium 自动化技术 +APScheduler 框架进行多时段网络资源数据获取 [3]。其次进行数据的去重清洗，使用 MySQL 数据库进行结构化存储，构建多源地震应急处置灾情信息语料。再次，针对地震应急处置灾情长文本信息中多类型事件实体重叠抽取不准确及语义复杂不明确的问题，提出一种基于机器阅读理解框架的地震应急处置信息事件抽取模型，引入先验知识编码结合双向 GRU 网络 [4]。然后采用基于触发词的多轮问答模式对事件论元抽取进行起始位置预测，从而较好解决了实体重叠问题，增强了信息文本的语义特征，提高事件论元抽取的精准度。最后使用 Element+Arcgis 三维地图和 JavaScript 开发方法在 Web 前端对抽取到的信息进行展示。具体技术流程图如图 1 所示。

2.2 震后地震应急处置信息语料库构建

在地震发生后，许多官方媒体和社交媒体会发布大量的新闻和一些政府相关处置信息，采用基于 Python 的 Selenium 自动化技术进行浏览器操作，使用相关 Windows 控件并结合高级调度器 APScheduler 模块多时段分布式任务调度获取地震应急信息。本文以近 10 年的历史灾情信息构建灾情信息语料库，对获取后的地震应急处置灾情文本信息制定一套数据采集规则，按照地震名称、发震时间、发震震级、地震应急处置信息内容、地震应急处置信息发布时间，地震应急处置信息发布来源进行结构化存储。根据云南省地震局白先富研究员等人发表的《地震应急现场信息分类初步研究》一文中的地震应急现场信息分类表 [5-6] 分为人员伤亡、救援出队、救援物资、道路损坏、房屋破坏 5 类地震应急处置灾情类事

件（如表 2），并对应论元角色进行指针序列标注。同时，针对地震应急处置信息语料库数据量较小的问题，通过问题 + 片段文本的方法对同一段落中的不同实体进行表示，对事件问句进行反义词替换、否定词添加等方法构造不可回答问题，不但将数据量扩大为原始采集数量的 53 ～ 80 倍，并且提高各类震后应急处置事件抽取模型的健壮性和鲁棒性，最终构建地震应急处置信息事件语料库。

表 2 地震应急处置类事件及对应角色表

序号	事件类型名称	触发词	事件论元角色名称
1	人员伤亡	受伤、死亡、轻伤、重伤等	时间、受伤人数、死亡人数、伤亡地点等
2	救援出队	派出、派遣、带、调派等	时间、出队单位、出队人数、负责人、出发地等
3	救援物资	调配、运送、携带、提供等	时间、物资数量、物资类型、物资提供者、物资发出地点、物资提供者、物资目标地点等
4	道路损坏	阻断、破坏、损坏、受阻等	时间、道路损坏地点等
5	房屋损坏	受损、损毁、损坏、破裂等	时间、房屋损坏地点等

3 关键技术与算法

本文提出基于机器阅读理解框架的地震应急处置信息事件抽取模型，以“问题 + 事件句”输入方式结合 BERT 编码获取特征向量。对于事件检测任务，引入标注好的触发词先验知识库，结合双向 GRU 获得全局特征向量，进行基于机器阅读理解框架 + 先验知识库的事件检测模型；对于事件论元抽取任务，采用基于触发词和问答模板（如表 2）的多轮问答模式，通过条件正则层融入触发词语义特征，对事件论元抽取进行起始位置预测，最终获得完整的结构化事件信息，如图 2 所示。首先，使用 BERT-WWM（Whole Word Mask）预训练模型学习到词的语义信息 [7]，整体模型的输入是震后应急处置信息问句与震后应急处置信息上下文的拼接结果，例如图 2，给定一段输入文本：“[CLS] 调配的时间是什么？[SEP] 截至 9 月 6 日 10 时，四川省应急管理厅紧急调配救灾帐篷 4400 顶 [SEP]”。其中，[CLS] 表示开始指示符，[SEP] 表示分割指示符。然后将拼接好的问答文本序列喂入模型中进行编码，在隐藏层进行输出，从而通过模型学习问题和文本的交互信息，最终得到融入了问题语义信息的地震应急处置信息事件文本向量。最后将训练好的 BERT 模型中的权重和偏置项作为条件正则层的输入条件，通过条件正则化的方法，得到融合了地震应急处置信息抽取后的触发词的语义信息的文本向量，从而提高了模型预测论元的准确率 [8]。利用双向门控循环神经网络（BiGRU）来学习合成的地震应急处置文本向量的整体特征，使用 BiGRU 模型可以更好地全面捕捉序列数据中的特征信息，并结合自注意力机制，为网络提供各个时间步不同的辅助信息，最后将两个方向的隐藏状态拼接后输出。其中，正向 GRU 表示融合后的地震应急处置信息词向量序列从 e^1_i 到 e^t_i 读取中的每一个 token，逆向 GRU 表示从 e^t_i 到 e^1_i 读取中的每一个 token，最后输出得到的学习到先验信息的隐向量表示 h^t_i [9]，公式为：

$$\overrightarrow{h_i^t} = \overrightarrow{GRU}\left(e_i^t\right), i \in (0, n), t \in (1, T) \quad (1)$$

$$\overleftarrow{h_i^t} = \overleftarrow{GRU}\left(e_i^t\right), i \in (0, n), t \in (1, T) \quad (2)$$

将双向 GRU 的输出结果输入论元输出层，并映射为一个二维矩阵。其中一维表示论元起点位置的置信度 start，另一维表示论元终点位置的置信度 end，最终得到地震应急处置事件论元起点位置和终点位置的矩阵，并通过多轮问答的方式对每个问句对应的答案进行事件论元预测输出（如图 2 所示）。

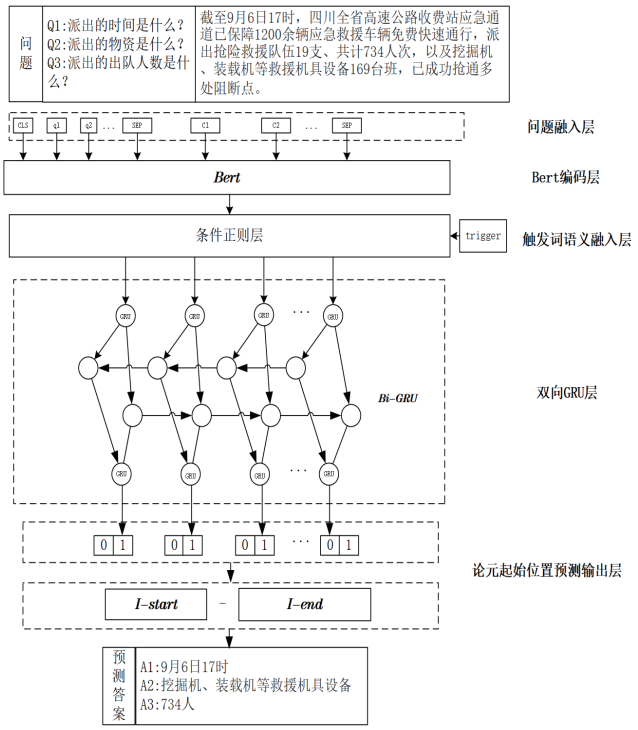

图 2 基于 MRC 的震后应急处置事件论元抽取模型

4 系统模块介绍

4.1 新闻事件管理模块

对使用了 Selenium 自动化库 +APScheduler 模块自动获取后的地震应急处置灾情新闻进行数据预处理，最后使用 Element 组件进行各个地震的地震应急处置灾情新闻事件查看，可以对事件新闻进行增加、删除、修改，同时按照日期、震级、震中位置进行检索操作。新闻事件管理模块如图 3 和图 4 所示。

图 3 新闻管理模块

图 4 事件管理模块

4.2 地震应急处置事件可视化模块

首先采用基于 Python 的 Selenium 自动化技术 +APScheduler 模块 + 基于阅读理解框架的双向 GRU 网络的地震应急处置灾情信息模型进行人员伤亡、救援物资、救援出队、道路损害、房屋损坏等标注点事件数据多时段获取自动化获取和事件抽取分析。然后使用 ArcGIS API 作为地理图层基本展示框架，将抽取后的各类灾情标注点信息在地图上映射添加显示。最后通过点击“事件”按钮展示对应事件列表，点击地图中的标注点事件进行抽取内容展示。图 5 ～图 7 为 2022 年 9 月 5 日四川甘孜州泸定县 6.8 级地震震后 6 h（出现人员伤亡信息 2 条、房屋损坏信息 2 条，如图 5 所示），震后 12 h（出现人员伤亡信息 5 条、房屋损坏信息 11 条，如图 6 所示），震后 24 h（出现人员伤亡 12 条、房屋损坏 21 条、道路损坏 8 条等关键应急处置灾情信息，如图 7）的动态标注事件展示图。

图 5 震后 6 h

图 6 震后 12 h

图 7 震后 24 h

5 结论

针对震后黑箱期灾情信息获取难度高、灾害汇聚多源繁杂的问题，本文采用基于 Python 的 Selenium 自动化技术 +APScheduler 定时模块进行数据多时段获取自动化获取，以近 10 年的历史灾情信息构建灾情信息语料库，并构建 MRC 机器阅读理解框架，对地震应急处置文本信息进行事件抽取，根据对应事件抽取模板，采用多轮问答模式引入触发词先验知识编码，结合双向 BiGRU 网络设计了基于网络资源的震后应急处置灾情信息可视化平台。但目前该方法时间开销大，泛化和鲁棒性有待增强，在未来可以考虑在问题构建的模式或者交互信息上尝试改进，以进一步提升抽取准确性。

参考文献：

[1] 王晓振 . 分布式浏览器自动化测试系统的设计与实现 [D]. 广州：华南理工大学 ,2014.

[2] 郭晓云 . 基于 Python 和 Selenium 的新浪微博数据访问 [J]. 电脑编程技巧与维护 , 2012(15):21-23.

[3] 张嘉威 , 关成斌 . 基于 Python 和 Selenium 的智联招聘数据的爬取与分析 [J]. 软件 ,2022,43(8):170-175.

[4] 葛君伟 , 乔蒙蒙 , 方义秋 . 基于上下文融合的文档级事件抽取方法 [J]. 计算机应用研究 ,2022,39(1):48-53.

[5] 白仙富 , 李永强 , 陈建华 , 等 . 地震应急现场信息分类初步研究 [J]. 地震研究 ,2010,33(1):111-118+120.

[6] 涂飞明 , 刘茂福 , 夏旭 , 等 . 基于 BERT 的阅读理解式标书文本信息抽取方法 [J]. 武汉大学学报 (理学版), 2022, 68(3): 313-316.

[7] 翟羽佳 , 许佳 , 李晓 . 面向突发重大公共卫生事件的多源异构应急信息融合模型研究 [J]. 图书与情报 ,2021(5):9-20.

[8] 吴旭 , 卞文强 , 颉夏青 , 等 . 机器阅读理解式中文事件抽取方法 [J]. 计算机工程与应用 ,2023,59(16):93-100.

[9] 安娜 , 白雄文 , 王红艳 , 等 . 基于双流注意力机制的阅读理解式事件抽取模型 [J]. 计算机工程与设计 , 2022, 43(6): 1686-1693.

【作者简介】

丁宸炀（1999—），女，湖南长沙人，硕士，研究方向：自然语言处理。

黄猛（1976—），男，河南新乡人，硕士，教授，研究方向：GIS、软件工程、机器学习、深度学习、大数据分析。

任玺睿（2004—），女，仡佬族，贵州铜仁人，学士，研究方向：计算机科学与技术。

邵文博（2004—），男，河南郑州人，学士，研究方向：计算机科学与技术。

（收稿日期：2023-11-17）