战“疫”记忆库构建关键技术研究

季士妍[1] 曾月清[2] 刘雅珺[3] 刘 耀[3]
（1.国家图书馆，北京 100034；2.北京大学，北京 100871；
3.中国科学技术信息研究所，北京 100038）

摘要：研究战“疫”记忆库构建的关键技术，不仅能为针对性解析各类战“疫”记忆资源提供技术手段与方法，还能将各类资源的知识结构关联起来，从而将零散的信息和知识组织起来，为自动化构建记忆库提供一整套技术与方法。针对人物与事件两类资源，提出战疫记忆人物专题构建流程与方法、事件与事件关系判定方法，实现自动生成战疫记忆人物专题、战疫事件因果与顺承关系的判定，并设计构建战“疫”记忆库展示平台，对各类资源进行统计、关联与多角度的可视化，验证研究技术和方法的可行性。通过对各部门和机构的相关信息进行收集和分析，将较为独立与分散的各部分资源形成关联化的资源体系，为专题记忆库的构建提供方法和技术方面的参考。

关键词：新冠肺炎疫情；构建记忆库；命名实体识别；事件提取；事件判定词表
DOI：10.3772/j.issn.1674−1544.2022.03.002 **CSTR：**15994.14.issn.1674−1544.2022.03.002
中图分类号：G250 **文献标识码：**A

Research on the Key Technology of Fighting COVID-19 Memory Base Construction

JI Shiyan[1], ZENG Yueqing[2], LIU Yajun[3], LIU Yao[3]
(1.National Library of China, Beijing 100034; 2.Peking University, Beijing 100871; 3.Institute of Scientific and Technical Information of China, Beijing 100038)

Abatract: This paper studies the key technologies of fighting COVID-19 memory base model, which can not only provide technical methods for the targeted analysis of various memory resources, but also can associate the knowledge structure of various resources, so as to organize scattered information and knowledge, and provide a set of technologies and methods for the automatic construction of memory base model. Targeting on characters and events resources, the paper proposes the process and method for the construction of fighting COVID-19 topic related to people, and the method of determining the relationship between events were proposed, which realized the automatic generation of fighting COVID-19 topic related to people and the extraction of causal and temporal relationships between fighting COVID-19 events. Finally, the research design and develop the display platform, and all kinds of resources are displayed in various aspects, to verify the feasibility of the research technology and method. Through the collection and analysis of relevant information of various departments and institutions,

作者简介：季士妍（1978—），女，硕士，国家图书馆副研究馆员，研究方向为数字资源知识化建设、数字资源网络采集及保存；**曾月清**（1994—），女，北京大学硕士生，研究方向为计算机辅助翻译；**刘雅珺**（1997—），女，中国科学技术信息研究所硕士生，研究方向为自然语言处理与人工智能；**刘耀**（1972—），男，中国科学技术信息研究所研究员，研究方向为自然语言处理、知识工程与知识发现（通信作者）。

基金项目：国家社会科学基金一般项目“数字资源知识共享与知识再利用模式与方法研究”（21BTQ011）；2020 年度文化和旅游研究项目“智能化时代的公共数字文化建设研究”（21DY28）。
收稿时间：2021 年 12 月 15 日。

this paper forms an associated resource system from relatively independent and dispersed resources, which can provide methodological and technical reference for the construction of thematic memory base model.
Keywords: COVID-19, construction of memory base, named entity recognition, event extraction, event decision vocabulary

0　引言

2019年年底爆发的新型冠状病毒肺炎疫情（简称“新冠肺炎疫情”），是人类共同面临的严峻危机，给社会带来了重大影响，全球抗疫的经历共同构成了人类应对公共卫生危机的集体记忆[1]。为了使得人类在未来能够针对新冠肺炎疫情进行全面的总结，分析新冠肺炎疫情对于人类社会发展的影响，联合国教科文组织文献遗产部发布了《转危为机——利用新型冠状病毒疫情为开展文献遗产工作争取更多支持》，倡议各组织各部门能够收集整理有关新冠肺炎疫情的各种信息和资源[2-3]。因此，我国众多机构和组织也在积极地对新冠肺炎疫情相关资源项目进行建设，收集各种国内外疫情相关的信息和资源。

然而新冠肺炎疫情的爆发和人类对疫情的抗击是一个动态的过程。在这一过程中，又分为了不同的阶段。在每个阶段各机构或组织都会收集和处理相关的数据或资源，而大部分的机构所获取的资源都相对比较分散和独立，由于没有统一的系统或者知识库导致各机构的资源无法相互关联，不能提供全方位的疫情记忆资源服务系统。同时，由于该疫情又属于突发性公共卫生危机，目前各部门在数字化技术收集、整理和存储相关报道、政策等资源时存在以下问题：一是多源异构资源的概念模型还未构建，无法自动解析资源结构和内容；二是当前大部分机构收集的资源比较分散和独立，无法形成关联化的资源体系，更无法构建能够串联记忆的战“疫”记忆库模型；三是还未形成相对成熟的相关的知识图谱技术，无法深入分析资源结构语义和资源内容语义；四是还未开发出具有展示战“疫”数据的平台，因此无法实现战“疫”信息和资源的可视化，给当下和未来的研究造成了一定的困难。以上问题都阻碍着人类对于新冠肺炎疫情认识的发展和总结。

人类已经进入了数字化新时代[4]，为了人类能够保护和传承对于共同应对新冠肺炎疫情这一公共卫生危机的集体记忆[5]，利用先进的技术手段将各种资源进行关联化和可视化[6]，实现对记忆资源的自动化构建就显得尤为重要。然而，当前实现这一过程的自动化程度较低，能够自动化构建记忆库来实现知识关联和知识组织的成功例子少之又少。本文研究了战“疫”记忆库的系统设计、模型构建流程和关键技术及方法，对于从各方面获取各类原始资源，形成相关数据集，构建各类型描述模型以及对于人物和事件专题的构建，完成了元数据集的内部关联以及元数据集与外部数据的关联建设[7]，从而完成了数字化、立体化战“疫”记忆库的构建。

1　战“疫”记忆库定义及研究意义

立足于当前的大数据环境，对于数字资源记忆的需求[8]、组织知识的标准以及对于知识管理的需求日益迫切[9]，战“疫”记忆库的构建需要完成语义化描述以及语义互操作，因此构建中国战“疫”记忆库需满足以下条件：一是具有基本的概念模型，使得领域知识能够形象地表达出来，形成比较完整的概念模型体系；二是明确化，能够将知识明确地定义出来，并且将隐性知识显性化；三是计算机可理解，将概念、概念属性以及概念关系能够用计算机可以理解的语言描述出来；四是关联，概念模型之间能够通过概念属性相互关联起来，使得记忆库能够实现一体化；五是共享，模型中的各概念应当是相关领域中的公共知识，并且能够进行公共知识管理。

本文通过构建战“疫”记忆库，将分布在各机构的零散的数据和信息整合在一起，并且对各

类资源进行解析和关联[10]，形成知识建设体系。为实现以上功能提供关键技术及方法，且为自动化构建整体战“疫”记忆库提供相关技术。

本文的研究具有以下意义：①建立各类资源的描述模型和记忆库构建模型。通过收集、整理和存储各机构和部门对新冠肺炎疫情的相关报道以及收集各文献信息等方式获取原始的数据资源，对原始资源进行解析，并统一对元数据规范化处理，这一过程解决了当前简单存储资源、未对数据进行规范化处理的问题，为后续将各类资源关联起来奠定了良好的基础，并且针对各类资源建立描述模型和战“疫”记忆库模型，统一数据规范，具有重要的意义。②提供战“疫”记忆库构建的关键技术和方法。目前利用自然语言处理技术解析战“疫”记忆资源的研究并不多，因此资源内容语义和资源结构语义分析不够深入，资源相对分散和独立。本文使用关键技术和方法从内容和结构两方面出发，提取[11]和加工战“疫”记忆库中涉及的数据、概念、关联关系、结构等资源，为实现战“疫”记忆库的自动构建奠定技术基础。③战“疫”记忆库的构建。本文对战“疫”记忆库进行构建和功能实现，提供战“疫”知识库知识服务，为用户提供有关新冠肺炎疫情各方面知识和资源，将人物与事件相互关联，使得用户能够更加全面清晰地了解抗击新冠肺炎疫情的记忆知识。

2 战“疫”记忆库系统设计

2.1 战“疫”记忆库模型及构建流程

战“疫”记忆库内容上主要涉及信息资源当中的概念、实体、事件以及各种关系，而最需要突破的难点则是从信息资源中识别抽取出相关的概念、实体和事件。本文在研究中界定的记忆库模型主要包括 3 个部分：战“疫”记忆概念模型、战“疫”记忆概念关联模型、战“疫”记忆概念属性取值模型。其中，记忆概念模型主要界定了相关概念及属性、记忆概念关联模型界定了概念之间的相关关系、记忆概念属性取值模型界定了各属性值的来源。战“疫”记忆库模型如图 1 所示。

针对业务需求分析和相关研究，在构建战“疫”记忆库时，首先利用爬虫和调用公共接口的方式从新冠肺炎疫情相关新闻、中文百度百科和期刊论文、学位论文和学术成果等方式获取原始信息资源，并通过构建原始资源描述模型对以上资源进行解析和存储；然后根据战“疫”记忆概念属性取值模型，通过实体和实体关系识别、事件和事件关系抽取[12]以及自然语言处理技术[13]

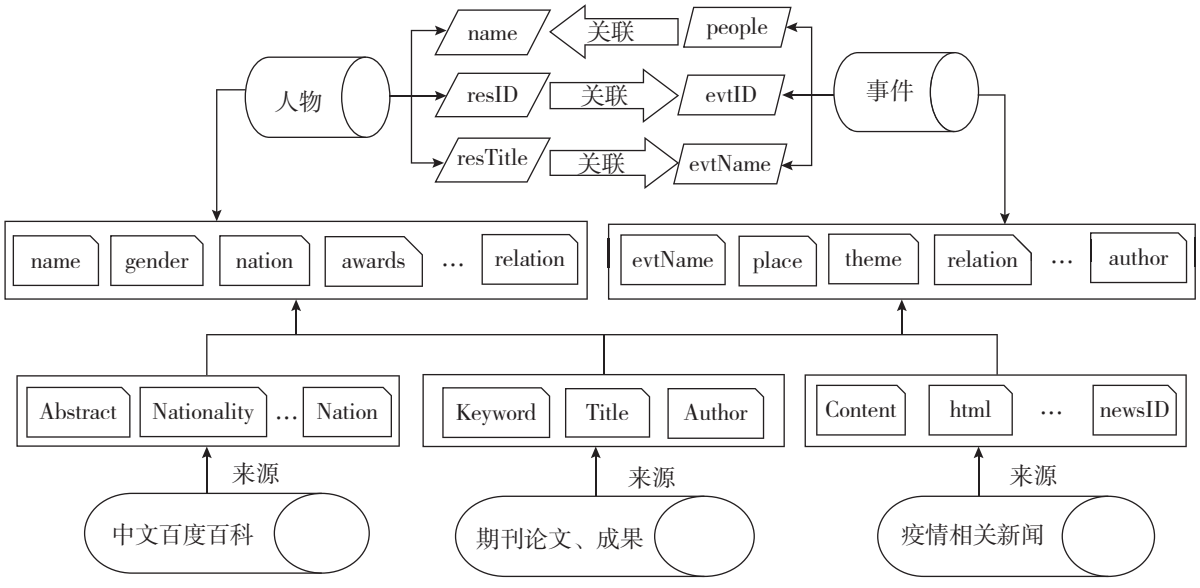

图 1 战“疫”记忆库模型

对原始资源进行解析处理，来获取资源中的实体和实体关系、事件和事件关系[14]；最后将处理好的资源按照战“疫”记忆概念模型和战“疫”记忆概念关联模型的定义进行组织关联[15]，形成战“疫”记忆资源网络体系，最终以一体化、立体化的结构和可视化的形式将战“疫”记忆完整展示出来[16]。记忆库构建流程如图 2 所示。

2.2　原始资源获取及模型构建

2.2.1　原始资源获取及其描述模型构建

战“疫”记忆原始资源包括半结构化数据资源和非结构化数据资源。其中，半结构化数据资源重要来源于中文百度百科、期刊论文、学术论文和成果，这些资源是通过调用万方数据的接口而获取的。非结构化数据具有完全或几乎无层级结构、无属性关联词描述的特点，非结构化数据资源主要来源于新闻网站上使用人工归纳的疫情关键词爬取新冠肺炎疫情相关新闻。另外，由于有关新冠肺炎疫情的新闻多且杂，在利用爬虫技术爬取新闻时，若获取全部内容，会给后续的数据清洗和概念识别工作造成一定的困难，因此在获取非结构化数据时，应当有选择地提取只用于概念和关系的部分数据。本文将从百度百科、万方数据、新浪新闻等平台获取相关数据，如表 1 所示。

由于结构化数据具有明确的数据类型、逻辑结构和相关关系，并且计算机能够对其进行理

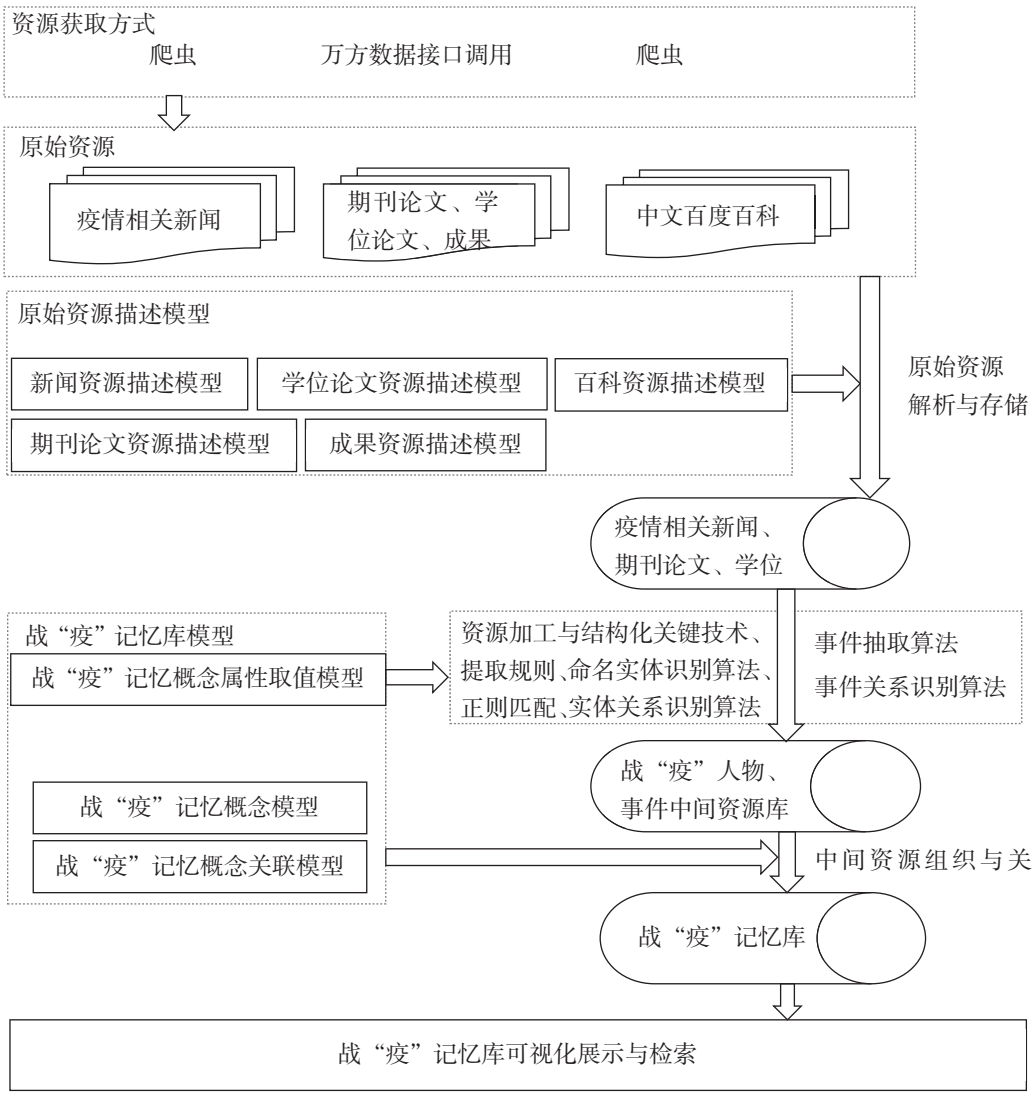

图 2　战“疫”记忆库的构建流程

表 1　数据来源平台列表

类型	来源	获取对象简介
半结构化数据	百度百科（baike.baidu.com）	开放百科平台人物词条
	万方数据（wanfangdata.com.cn）	成果、期刊论文、学位论文
非结构化数据	新浪新闻（news.sina.com.cn）	新冠肺炎相关新闻

解和处理，因此结构化数据是最理想的信息表达形式。然而，结构化数据很少出现在现实的世界中。为了更加方便地抽取战“疫”记忆相关概念及其关系，需要构建战“疫”记忆原始资源描述模型以此来结构化解析资源。资源描述模型会对元素、属性、结构和数据类型等进行定义，数据解析工具会根据其定义对爬取的数据进行过滤、名称标准化和统一XML格式转换，以便于生产和复用各类文档资源和数据库资源。因此，构建资源描述模型非常重要。下面分析各类资源并根据分析结果构建其相应的资源描述模型。

（1）人物百科资源描述模型，是基于人物百度百科资源结构进行的。词条主要包括人物简介以及中文名、外文名、国籍、民族、出生地、出生日期等基本信息，并基于此构建人物百科资源描述模型。

（2）新闻资源描述模型，一般包括新闻的基本信息、新闻文本内容、新闻爬取的元数据信息。基本信息包括标题、发布时间、来源、作者及新闻类型。元数据信息主要包括新闻的爬取时间、html网页快照、关键词等。基于此构建新闻资源描述模型。

（3）期刊论文资源描述模型，主要是基于万方数据的期刊论文资源构建的。期刊论文资源包括篇名、英文篇名、论文基本信息、正文、引文等元素。其中，论文基本信息包含作者、作者单位、中文摘要、中文关键词、英文摘要等多个元素。基于此构建期刊论文资源描述模型。

（4）学位论文资源描述模型，主要是基于万方数据的学位论文结构资源构建的。学位论文资源包括标题、论文基本信息、正文、引等元素。其中，论文基本信息包含作者、学科专业、授予学位、学位授予单位、导师姓名等。基于此构建学位论文资源描述模型。

（5）成果资源描述模型，是基于万方数据的成果结构资源构建的。成果资源包括标题、项目年度编号、完成单位、完成人、成果公布年份、中图分类号、关键词和成果简介等，并根据这些元素构建成果资源描述模型。

2.2.2　战“疫”记忆库模型构建

战“疫”记忆库模型主要包含战“疫”记忆概念模型、战“疫”记忆概念关联模型和战“疫”记忆概念属性取值模型。

（1）战“疫”记忆概念模型。主要界定了战“疫”记忆所涉及的概念及其属性。本文研究只有借助Schema校验工具将非结构化数据和半结构化数据转化为符合要求的结构化信息，并进行解析之后，才能得到进一步结构化的数据，为后续工作奠定基础。本文构建了战“疫”人物和战“疫”事件两种概念模型。其中，战“疫”人物概念模型的知识单元为“人物”，并以“人物”为核心进行描述，包括姓名、性别、国籍、民族、图像、出生和死亡时间、出生地、籍贯、职业、工作机构、贡献成就、相关资源、人物关系等，构建人物Schema，并以<people>作为根节点，方便后续抽取资源标签下的内容。战“疫”事件模型的知识单元为“事件”，并以“事件”为核心进行描述，包括事件信息、事件描述、相关事件、事件涉及的新闻元数据4个部分。事件信息包含事件名称、开始时间、相关地点、相关组织机构、相关人物、事件分类；事件描述为涉及该事件的相关段落和句子；相关事件包含相关事件名称、相关事件ID、事件关系；事件涉及的元数据包含事件来源网站、来源ID、资源类型、发布时间、作者、关键词。事件Schema根节点为<COVID19Event>，包括事件信息、事件描述、相关事件、元数据等自定义元素，各二级元素又包含多个三级元素。

（2）战“疫”记忆概念关联模型。战“疫”记忆人物和事件模型之间能够通过一个或者多个概念的共有属性相互关联，并且该记忆库模型内部元素同样相互关联，通过人物关系和事件关系分别与其概念模型本身相关联。表2展示了战“疫”记忆人物和事件概念的属性关联。

（3）战“疫”记忆概念属性取值模型。主要是以战“疫”人物概念模型和战“疫”事件概念模型的属性为核心，从原始资源描述模型中获取相应的值。

3　战“疫”记忆人物专题构建技术和方法

3.1　战“疫”新闻语料预处理

首先，过滤非事实描述类新闻。本文主要基于所爬取的疫情新闻语料[17]来识别实体并判定人物关系，因此可以基于标题特征、内容特征对新闻语料进行过滤。本文是基于事实描述类新闻对事件进行提取的，将不符合要求的疫情新闻标题通过正则表达式进行过滤。过滤与疫情无关的人物，将爬取的文本中不相关人名，如记者、摄影等，这些短语具有明显的符号特征，使用正则表达式进行过滤，而过滤特征不明显的词语，可以使用词表结合位置信息的人名标注过滤方法。

然后，对通用命名实体识别工具适用性进行研究。利用自然语言工具对疫情新闻语料进行预处理。疫情相关新闻命名实体主要为地点、人物和组织机构，因此本文选用了常用命名实体识别工具——HanLP、哈尔滨工业大学中文依存句法分析工具（LTP）的命名实体识别功能对疫情相关新闻语料进行命名实体识别[18]。两者在人名和地名识别上相差不大，较能准确地识别出人名，但是对于地名两者均存在误差。在结构识别中，由于部分机构名字存在嵌套的情况，两者均未能准确识别出国家卫健委高级别专家组、浙江大学医学院附属第一医院传染病诊治国家重点实验室等机构名字。

3.2　战“疫”记忆命名实体识别技术

通过对疫情新闻文本进行分词，基于战“疫”记忆库构建的需求，需要识别出与疫情相关的人名、地名、机构组织名等命名实体，因此本文使用的技术为Bi-LSTM-CRF模型[19]和BERT-Bi-LSTM-CRF模型，对以上实体进行识别[20]，并增加了对事件发生时间的识别[21]。

首先，利用语料库对模型进行训练，对疫情发生的事件、地点、人物和组织机构进行识别；然后，计算出对实体进行自动识别的精确率（Precision）、召回率（Recall）和F1的值。其中，BERT-Bi-LSTM-CRF模型在识别疫情命名实体时效果最好，尤其是人物、时间和地点，但在识别组织机构时效果比较差，这可能是部分组织机构存在实体嵌套和缩略词的情况。

3.3　战“疫”人物关系识别与判定技术

战“疫”记忆库同样也需识别和判定人物关系。人物关系主要有学生、父母、师承、子女、朋友、同事、合作伙伴、配偶等。实体关系抽取过程是信息抽取的关键[22]，因此在人物关系自动识别中，首先要去除与疫情不相关人名的句子，然后将带有特殊符号实体标记的语料以8:1:1的比例划分为训练集、测试集和验证集，对R-BERT[23]模型进行训练。利用训练好的模型识别疫情新闻人物，输出人物关系。R-BERT模型[24]对于父母、朋友和子女的关系识别效果较好。但是当句子中没有明显表示关系的词语出现时，会出现关系识别错误的情况。

3.4　战“疫”人物专题构建方法

在对人物进行构建时，需填充人物属性并生成专题。采取的主要方法是：战“疫”记忆人物

表2　战“疫”记忆人物和事件概念的属性关联表

属性字段所属概念模型	概念字段名称	概念字段含义	关联属性名称	关联属性含义	关联属性来源概念模型
战“疫”事件	people	相关人物	Name	姓名	战“疫”人物
战“疫”人物	resID	资源ID	evtID	项目事件ID	战“疫”事件
	resTitle	资源标题	evtName	事件名称	战“疫”事件

专题主要围绕战“疫”人物概念模型进行构建，依据战“疫”人物概念模型Schema，利用规则，实体识别技术[25]和人物关系判定技术在疫情新闻文本中对人物属性内容进行提取[26]，并利用字段值构建检索式，在人物百科、期刊论文、学位论文和成果原始资源库中检索人物及其属性进行填充，形成战“疫”人物简单专题。基于战“疫”人物和战“疫”事件关联模型与战“疫”事件简单专题进行关联，形成战“疫”记忆人物专题。战“疫”记忆人物专题构建流程如图3所示。

4 战“疫”事件抽取与关系判定技术和方法

4.1 战“疫”事件识别与抽取

首先要对战“疫”记忆事件进行识别和抽取，主要分为对疫情事件句的识别和对疫情事件要素的抽取。对疫情事件抽取的思路是先识别疫情新闻文本中的符合定义的事件句，再根据事件形式化定义[27]，进行事件要素识别抽取，形成疫情事件的形式化表达[28]。本文将战“疫”事件的表示形式定义为五元组，包含事件发生时间、施事主体、对象、事件类型和事件触发词。本文的事件抽取方法有3个部分，即疫情事件判定词表、疫情事件句判定规则和基于依存句法分析及角色标注的事件要素识别模型。

事件判定词表包括事件类型表、子事件领域词表及事件触发词表。其中，事件类型表的构建方法主要是通过LDA模型和TF−IDF特征词—权值模型对疫情新闻文本进行主题挖掘，并进行人工归纳，总结了38种常见的疫情事件类型[29]。子事件关键词表的构建方法主要为TF−IDF模型识别及人工筛选后的关键词。事件触发词表的构建方法是通过人工总结的句法分析模板获取的高频触发词以及基于高频触发词扩展的近义词和关联词。

本文事件抽取任务主要有事件句的识别和事件抽取及要素角色判定。其中，事件句识别方法为：将预处理后的战役文本进行分句和分词，通过触发词表和领域词表，将候选事件句识别和筛选出来，形成各事件类别的候选事件句。事件抽取及要素角色判定方法为：疫情新闻文本经过自

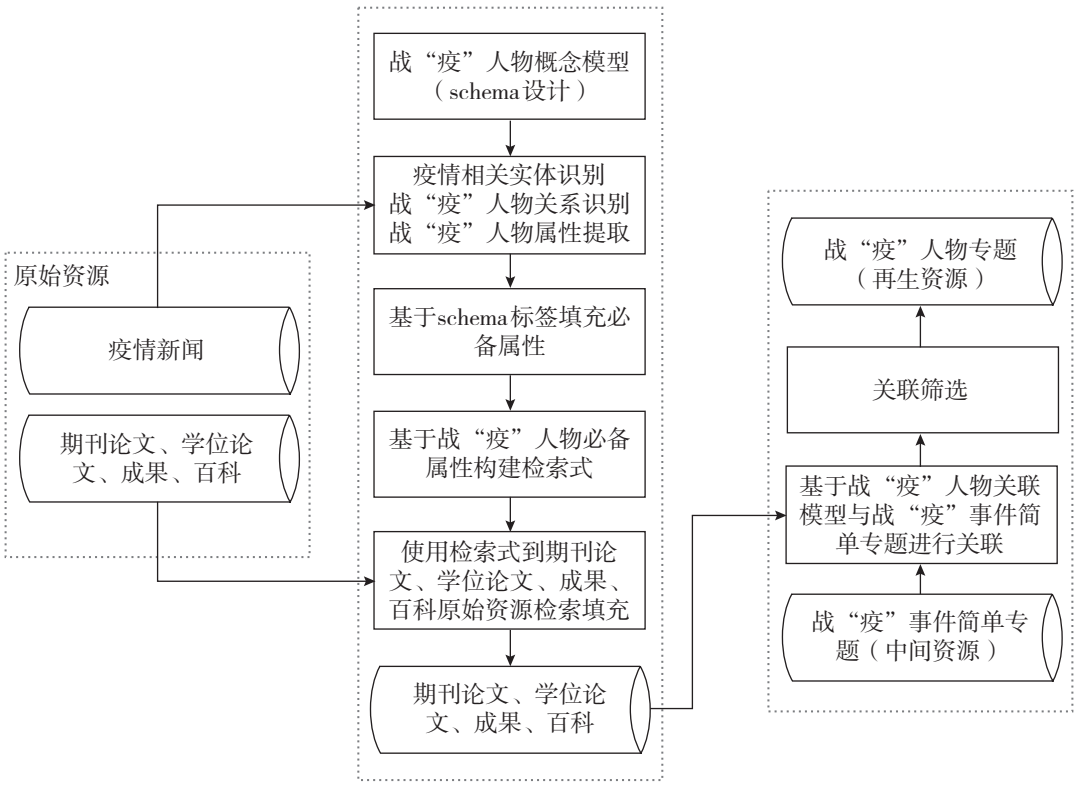

图3 战“疫”记忆人物专题生成流程

然语言处理工具的基本预处理后，便可将其输入疫情事件抽取模型，根据事件句判定规则识别事件句，然后基于依存句法分析和语义角色标注对抽取事件要素[30]。战“疫”事件抽取方法如图4所示。

4.2　战“疫”事件关系识别与判定方法

依据不同的语言特点，战“疫”事件关系主要有因果关系和顺承关系两类。

（1）因果关系的识别与判断。通过分析可知，具有因果关系的原因各结果事件通常在一个句子中。因此，对因果关系的识别与判定方法为：使用模式匹配的方式对因果关系进行识别和抽取，利用因果触发词构建匹配模板对每一个句子进行正则匹配，以此得到原因子句和结果子句，再将其进行抽取，便可得到原因事件和结果事件。战“疫”事件因果关系判定模型如图5

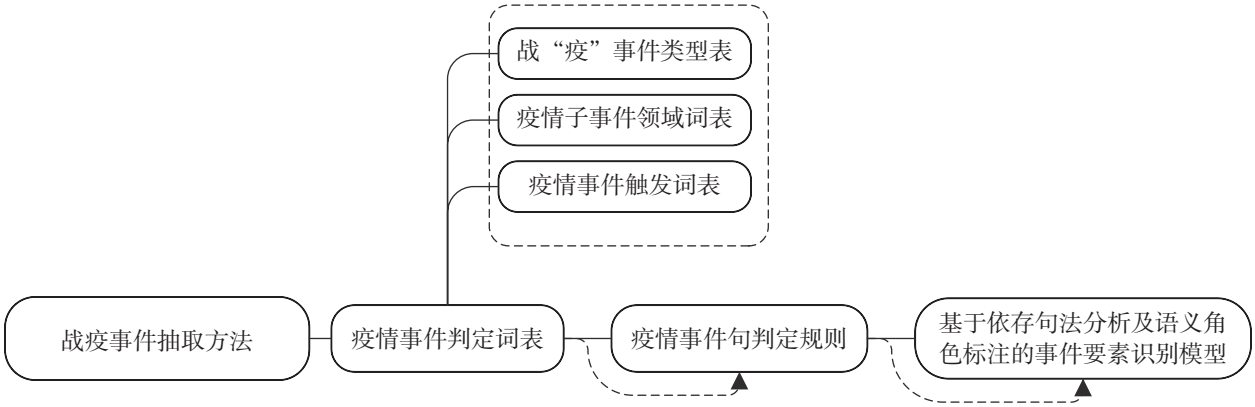

图4　战“疫”事件抽取方法

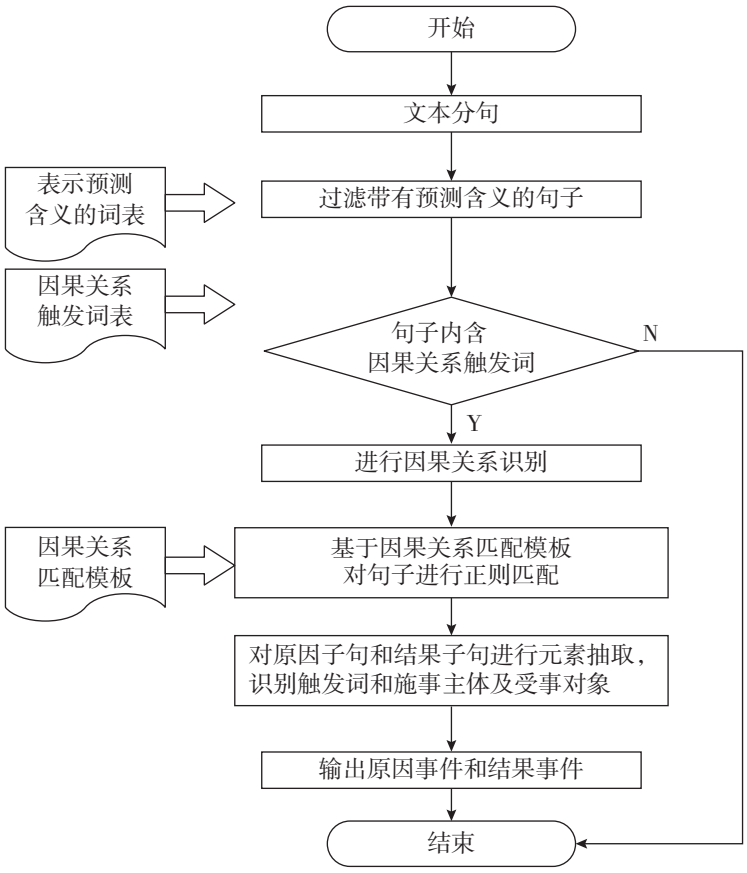

图5　战“疫”事件因果关系判定模型

所示。

（2）顺承关系的识别与判定。战“疫”记忆事件顺承关系没有明显的关系指示词，因此顺承关系的识别与判定主要是依靠语义关系进行判定，根据语义依存分析对战“疫”事件顺承关系进行判定和识别。首先对疫情候选事件句进行事件抽取，再利用LTP工具对疫情候选事件句进行语义依存分析，对事件句中词语的语义依存关系进行标记，进而找到满足后继关系的核心谓词对，然后将事件触发词和核心谓词对相互匹配，将满足的事件组成顺承关系事件对，最后根据语义依存分析的结果，判断事件对核心谓词索引值大小。值较小的为前序事件，值较大的为后继事件。战“疫”事件顺承关系判定流程如图6所示。

5 系统功能框架

5.1 功能展示

战”疫“记忆库专题框架如图7所示。本节将对战“疫”记忆库的抗疫新闻、抗疫人物、抗疫机构、学术成果和抗疫事件5个功能模块进行展示。

“抗疫新闻”功能模块展示了每日抗疫类的新闻，点击某一条新闻就可以进入该新闻的详细

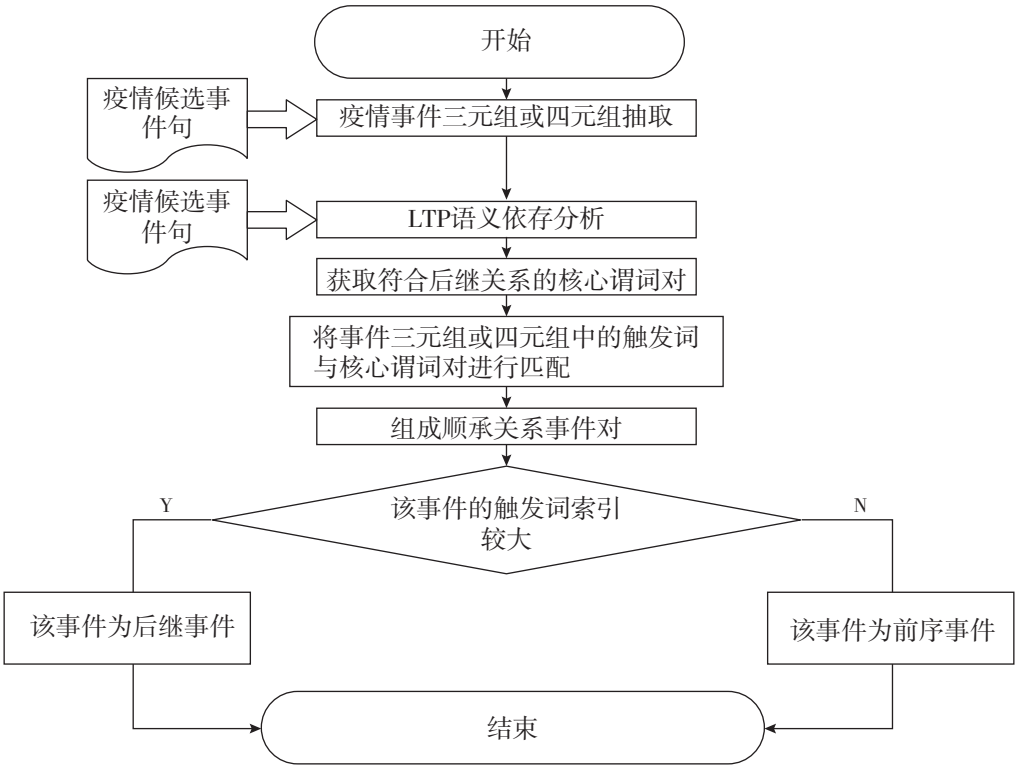

图6 战“疫”事件顺承关系判定模型

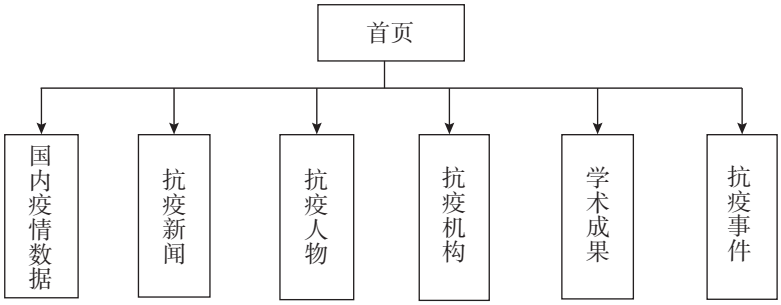

图7 战“疫”记忆库专题框架

页面。该页面分类展示了发布时间、相关人物、机构、地点和内容等。当进入抗疫新闻列表时不仅可以看到最新的抗疫新闻还可以对新闻进行检索，在详情页面中有原新闻的链接以及新闻的详细内容。同时，当新闻中有抗疫事件时，系统将会进行抽取和解析，判断出事件类型、时间地点以及机构等。其功能框架如图 8 所示。

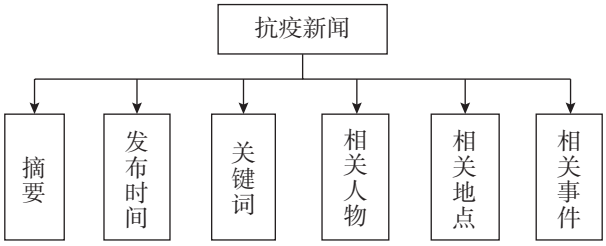

图 8　抗疫新闻功能框架

抗疫人物功能模块对抗疫人物进行了轮播展示。直接点击人物便可进入人物详情页面，该页面展示了人物基本信息、学术成果和相关抗疫新闻，并且通过可视化的方式展示出该人物的研究主题、人物关系、学术成果、相关事件等，可视化部分包括人物信息总览、人物研究主题、各种人物关系图谱和时间关系树图谱等，点击可以进入抗疫人物列表。其功能框架如图 9 所示。

抗疫机构功能模块对抗疫机构进行了轮播展示。直接点击机构便可进入机构详情页面，该页面展示了机构基本信息，是可视化展示。基本信息包括名称、外文名、简称、地址、所属部门和简介等信息；可视化部分包括相关事件、合作伙伴、研究主题和学术产出等。其功能框架如图 10 所示。

学术成果功能模块是对学术成果的展示，在学术成果列表页面可以对成果进行检索，也可查看期刊论文详情，其中包含期刊论文的基本信息、全文和摘要。其功能框架如图 11 所示。

抗疫事件功能模块可以对抗疫事件、新闻和人物进行检索。其中，在抗疫事件详情中，包括事件新闻的详细信息和源地址，以及事件的具体信息和事件句，提取出事件的时间、地点、人物和机构等。其功能框架如图 12 所示。

5.2　战“疫”记忆库测试及分析

本节将对战“疫”记忆库的人物板块功能和事件板块功能进行抽检测试，并对出现的错误和问题进行分析。

（1）在对人物板块功能测试后，通过分析抽检结果可知，存在识别出不相关人物如编辑、摄

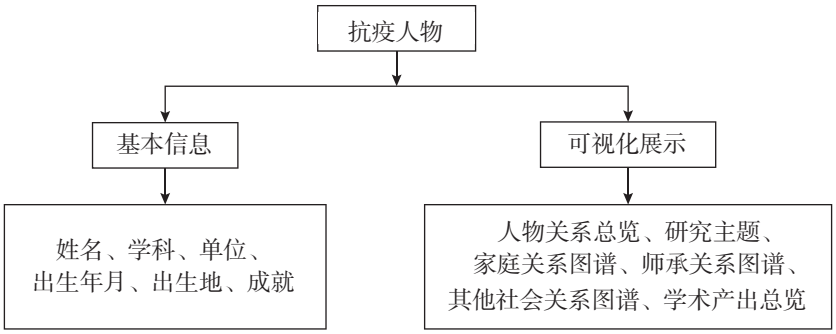

图 9　抗疫人物功能框架

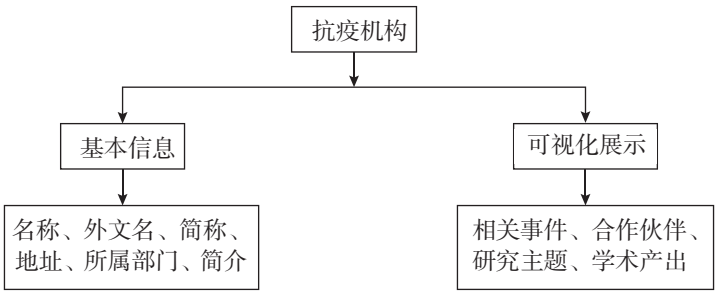

图 10　抗疫机构功能框架

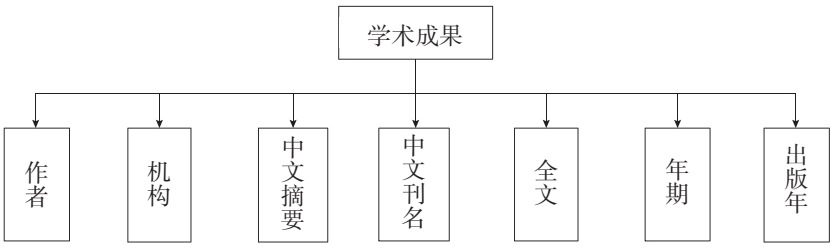

图 11 学术成果功能框架

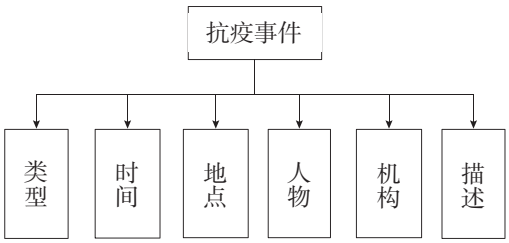

图 12　抗疫事件功能框架

影记者等的问题，还存在误将其他事物的名称识别成人名的问题，因此需要后续总结添加不相关人名的列表进行过滤。由于有些人物关系描述不存在显性关系标志词，如“妻子”“儿子”等词语，在进行人物识别时也会出现错误，造成误差。

（2）在对事件板块功能测试后，通过分析抽检结果可知，由于并不是所有的疫情新闻文本都带有时间关键词，因此存在对事件发生的时间识别不准确的问题。另外，子事件领域词表和事件触发词表数量有限，并且有可能部分词汇存在于多个事件类，导致一个事件被识别为多个事件类并且错误判定事件类型。施事者和受事者词汇包含嵌套实体和多个定语，导致两者之间的识别不准确。

分析抽检结果，总结出现的问题，并在关键技术和方法中对出现的问题进行相应的优化处理。

6　结语

本文对构建战“疫”记忆库模型进行了理论和关键技术的研究，根据结构化元数据集提供的信息构建了人物概念模型和事件概念模型，将获取的原始资源数据规范化、结构化入库。通过BERT−Bi−LSTM−CRF深度学习模型实现命名实体识别[31]以及应用事件抽取技术，实现了实体的自动发现和关系识别。通过分析战“疫”记忆资源的特点，对包含了各种资源实体内容和结构以及相互关联关系的战“疫”记忆库模型进行了构建。对各种来源的战“疫”资源进行描述，构建了资源描述框架，界定了记忆库模型涉及的信息字段以及各资源实体间的关系。本文采用一系列关键技术和方法，包括实体识别和判定[32]、人物关系的识别和判定、事件以及事件关系的识别与判定等技术与方法，优化了记忆库内容与结构的构建，验证了构建技术与方法的可行性。从多方位、一体化、立体化呈现了抗击疫情的记忆，将数据可视化、全方面地展示给用户。

但在解析战“疫”记忆资源的过程中，使用自然语言处理技术与深度学习技术挖掘资源的信息，大部分都是根据命名实体任务及事件抽取任务各自使用不同的方法进行的。此外，由于新冠肺炎疫情属于新近的突发公共卫生事件，该领域未有充足的标注语料，故未能训练与该领域内容相关的深度学习模型。今后的研究可以进一步探讨生成战“疫”记忆库的一体化算法，直接从各种战“疫”资源中生成战“疫”记忆库。

参考文献

[1] 加小双,姚静.国外高校“疫情记忆”实践的分析与启示[J].浙江档案,2020(8):16−18.

[2] Turning the threat of COVID−19 into an opportunity for greater support to documentary heritage[EB/OL].[2020−04−03].https://www.ica.org/en/unesco−statement−turning−the−threat−of−covid−19−into−an−opportunity−for−greater−support−to.

[3] 阿斯曼.文化记忆:早期高级文化中的文字、回忆和政治身份[M].金寿福,黄晓晨,译.北京:北京大学出版社,2015.
[4] 冯惠玲.数字记忆:文化记忆的数字宫殿[J].中国图书馆学报,2020,46(3):4−16.
[5] 蔡娜.重大事件档案管理机制研究[D].北京:中国人民大学, 2011:16.
[6] 周耀林,刘晗.数字记忆建构:缘起、理论与方法[J].山东社会科学,2020(8):50−59.
[7] 牛力,刘慧琳,曾静怡,等.数字时代档案资源开发利用的重新审视[J].档案学研究,2019(5):67−71.
[8] 冯惠玲. 数字时代的记忆风景[N]. 中国档案报,2015−11−19(3).
[9] 加小双,徐拥军.国内外记忆实践的发展现状及趋势研究[J].图书情报知识,2019(1):60−66.
[10] 周耀林,刘梦颖,杨文睿,等.后疫情时代抗疫数字档案资源整合[J].档案管理,2020(6):33−35.
[11] CHEN J X, JI D H, TAN C L, et al. Relation extraction using label propagation based semi−supervised learning[C]//Proceedings of the 21st International Conference on Computational Linguistics/44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Assoc Computational Linguistics−Acl, 2006.
[12] KONSTANTINOVA N. Review of relation extraction methods: What is new out there?[C]// International Conference on Analysis of Images, Social Networks and Texts. Cham: Springer, 2014.
[13] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre−training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[14] 唐敏.基于深度学习的中文实体关系抽取方法研究[D].成都:西南交通大学, 2018.
[15] 台丽婷.基于半监督机器学习的实体关系抽取算法研究[D].北京:北京邮电大学, 2018.
[16] 孔兵.中文文本实体关系抽取方法研究[D].哈尔滨:哈尔滨工业大学, 2016.
[17] YU S W, DUAN H M, WU Y F. Corpus of multi−level processing for modern Chinese[EB/OL]. Peking University Open Research Data Platform Dataverse, 2018 [2021−11−02]. http://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/SEYRX5.
[18] 李明扬,孔芳.融入自注意力机制的社交媒体命名实体识别[J].清华大学学报(自然科学版), 2019, 59(6):461−467.
[19] 曾勇.基于BiLSTM−CRF模型的中文命名实体识别研究与实现[D].南昌:江西财经大学,2020.
[20] HUANG Z, XU W, YU K. Bidirectional LSTM−CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
[21] GRISHMAN R. The NYU system for MUC−6 or where's the syntax?[R]. New York: New York Univ Ny Dept of Computer Science, 1995.
[22] 王传栋,徐娇,张永.实体关系抽取综述[J].计算机工程与应用,2020,56(12):25−36.
[23] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder−decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.
[24] WU S C, HE Y F, ACM. Enriching pre−trained language model with entity information for relation classification[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM). New York: Assoc Computing Machinery, 2019.
[25] 冯惠玲. 数字时代的记忆风景[N]. 中国档案报,2015−11−19(3).
[26] DODDINGTON G, MITCHELL A, PRZYBOCKI M, et al. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation[C]. Lisbon: European Language Resources Association (ELRA),2004.
[27] CHAKRABORTY S. Big data analytics for development: Events, knowledge graphs and predictive models[D]. New York: New York University, 2015.
[28] 吴超. 面向突发事件领域的事理图谱平台的设计与实现[D]. 成都:电子科技大学,2020.
[29] ANDRADE E L, BLUNSDEN S, FISHER R B. Modelling Crowd Scenes for Event Detection[C]// 18th International Conference on Pattern Recognition (ICPR 2006). Los Alamitos: IEEE Computer Society, 2006.
[30] 薛聪,高能,查达仁,等.事件库构建技术综述[J].信息安全学报,2019,4(2):83−106.
[31] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.
[32] LI J, SUN A, HAN J, et al. A Survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 99: 50−70.