www.csdata.org

ISSN 2096-2223
CN 11-6035/N

基于开源军事新闻的领域事件数据集

黄宏斌[1*]，孙皎[1]，韦晖[1]，肖开明[1]，王懋[1]，李璇[1]

1. 国防科技大学，大数据与决策实验室，长沙 410073

摘要： 军事领域的文本数据集是在军事领域进行事件抽取的基础，高质量的数据集可以有效地推动军事领域事件抽取的研究，然而目前常用的事件抽取数据集(如ACE2005等）是面向通用领域的，关于军事领域事件的文本语料资源十分匮乏。因此，我们从公开军事新闻网站获取大量的军事新闻内容；在对新闻文本内容进行分析的基础上，首先建立包含事件类型、实体类型和实体关系类型的军事新闻事件模型；进而依据事件模型对文本数据进行人工标注，同时对事件模型进行迭代验证和修正；最终获得由13000条高质量军事新闻文本事件组成的标注种类全、质量高的军事新闻事件数据集。本研究公开了该军事新闻事件数据集。

关键词： 自然语言处理；事件抽取；军事新闻；数据集

数据库（集）基本信息简介

数据库（集）名称	基于开源军事新闻的领域事件数据集
数据通信作者	黄宏斌（hbhuang@nudt.edu.cn）
数据作者	黄宏斌、孙皎、韦晖、肖开明、王懋、李璇
数据时间范围	2010–2022年
数据量	25.6 MB
数据格式	*.json
数据服务系统网址	http://dx.doi.org/10.57760/sciencedb.j00001.00486
数据库（集）组成	数据集包含13000条军事新闻，以*.json格式存储，数据量为25.6 MB。

引　言

随着信息技术高速发展，网络已经成为信息传播的重要载体。事件作为信息的一种表现形式，是对现实世界所发生的事情的描述，将特定的人、物在特定时间和特定地点相互作用的客观事实通过自然语言表达出来[1]。事件抽取是从非结构化的自然语言文本中抽取出事件要素，以结构化形式表达的一项技术。事件抽取在网络舆情监控、突发事件告警、情报收集领域有着重要应用[2]，例如，当今网络上存储着海量新闻文本信息，事件抽取技术可以帮助决策者从大量的低价值情报中自动获取事件信息，减轻情报收集的工作量。

我国的军队建设已经进入信息化时代，将信息抽取技术应用到军事领域，可以通过事件抽取和事件关系抽取构建事理图谱，进而可以对军事事件进行预测。对军事新闻进行抽取，可以获取各方军队的最新动向，一定程度预知对方接下来的行为；对军事手册进行抽取，可以建立军事行动的电子指南，方便对手册进行

文献 CSTR：
32001.14. 11-6035.csd.2022.0072.zh
文献 DOI：
10.11922/11-6035.csd.2022.0072.zh
数据 DOI：
10.57760/sciencedb.j00001.00486
文献分类：信息科学

收稿日期：2022-09-03
开放同评：2022-09-14
录用日期：2023-01-16
发表日期：2023-03-10

* 论文通信作者
黄宏斌：hbhuang@nudt.edu.cn

检索教学；对过往战争案例进行抽取，可以对以前的作战行动进行概括，方便人们从过往战争中总结经验[3]。

对于通用领域事件，目前已有 ACE2005[4]、MAVEN[5]等成熟开源数据集，其推动了通用领域事件检测和事件抽取的研究[6-9]。其中，ACE2005 数据集是常用的事件抽取数据集，定义的事件类型共 8 大类，加上各次类共 33 小类。然而，ACE2005、MAVEN[4-5,10]等数据集面向通用领域，其定义的事件类型不能很好地直接应用于军事领域；同时，当前面向军事领域的事件抽取公开数据集较少，在一定程度上制约了军事领域事件抽取技术的发展。此外，目前尚未有统一明确的军事事件模型，从近年来发表的文献中发现，以往研究者构建的数据集包含的事件类型较少，标注种类不够完善。基于公开的和爬取自特定网站的语料，游飞以飞机和舰船作为事件主体，将军事事件类型分为了属性类事件、转移类事件和行动类事件 3 种类型，明确了每类事件对应的事件元素[11]。高鸿博首先从头条百科中爬取军事装备相关词条内容，经整理后形成语料。他将军事事件分为参战事件、作战事件、移动事件和部署事件 4 种事件类型，同时定义装备、时间、地点等 7 种实体作为事件元素[3]。战立莹依据战争系统论概念、军事用语概念两个维度对军事事件类型进行界定，确定为侦查、指挥、部署、机动、控制、协同、行动、保障、攻击、防守、封锁 11 类[12]。

本研究从公开军事新闻网站（如新浪军事、凤凰军事和网易军事等）收集大量军事新闻文本语料人工构建语料库，这在一定程度上能够保证语料库构建的质量。为了减少反复检查的工作量，本文采用了迭代标注的方法，根据初步制定的军事事件模型对语料进行标注，并在标注过程中根据语料标注情况和军事事件模型的相互反馈，迭代更新语料标注结果和军事事件模型，构建了一个标注种类全、质量高的军事新闻事件数据集。

1 数据采集和处理方法

首先，从新浪军事、凤凰军事和网易军事等公开军事新闻网站，爬取军事新闻文本。在数据清洗之后，将新闻文本进行分词。对于分词之后的新闻文本，进行文本分析，确定事件类型、实体类型以及关系类型。由于语料的标注需要符合军事事件模型的定义，军事事件模型也要契合语料的内容，而预先构建军事事件模型对语料进行标注（自顶向下）的方法或根据语料的内容构建军事事件模型（自底向上）的方法都不能很好地满足要求，因此本文拟基于迭代更新的方法同时构建军事事件模型及中文语料库。数据采集和处理过程如图 1 所示。

1.1 数据采集方法

从新浪军事、凤凰军事和网易军事三家新闻媒体网站，按照时间顺序，采用 Request 和 BeautifulSoup 方法抓取了从 2010 年 8 月到 2022 年 2 月共 40000 余条新闻。

1.2 数据预处理方法

首先，将爬取到的原始语料采用正则表达式对文本内容进行数据清洗，过滤缺失样本和无关信息，将清洗之后的文本按长度进行排序，最终保留 13000 条表达精炼的高质量新闻文本。其次，对经过数据清洗后的文本使用哈工大 LTP 分词工具[13]进行分词操作，为了保证一些军事领域专有名词分词结果的正确性，本研究预先搜集军事领域名词加入分词词典，以改善分词结果。

1.3 军事新闻文本数据集的构建方法

当前面向军事领域的事件抽取公开数据集较少，本研究参考相关公开数据集和文献，分析爬取到的新闻文本内容，进一步确定了军事新闻文本数据集的事件类型、实体类型和关系类型，采用人工标注方式进行数据集标注。

首先定义了试验、演习、部署、保障、支援、意外事故、非事件等 7 种军事事件类型。在标注过程中，发现有相当一部分文本体现了“[时间]，[主体]在[地点]展示[装备]”，因此新增展示事件类型。8 种事件类型如表 1 所示。

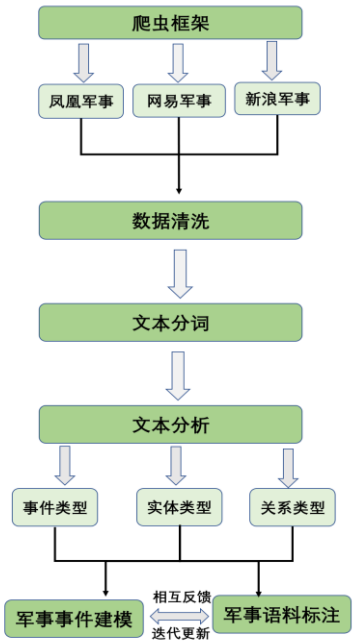

图 1 军事新闻文本数据采集和处理过程

Figure 1 Collection and processing process of military news text data

表 1 军事事件类型及对应事件元素

Table 1 Military event types and the corresponding event elements

事件类型	事件元素
试验	主体、装备、时间、地点
演习	主体、演习内容、时间、区域
部署	主体、军事力量、时间、地点
保障	主体、客体、物资、时间
支援	主体、客体、时间
意外事故	主体、事故后果、时间、地点
展示	主体、装备、时间、地点
非事件	——

为充分利用文本中的信息，本研究还定义了人物、设施、组织和装备等 7 种实体类型（如表 2 所示）及隶属、共事、合作和敌对等 7 种实体关系（如表 3 所示）。在标注过程中，发现文本中通常用不完全相同的词汇表达同一实体，如“GAR-9 空对空导弹”和“GAR-9 导弹”“苏弗朗号”和

“苏弗朗号潜艇”。因此新增等价关系，连接相同实体。由于等价关系的存在，多次出现的实体与其他实体之间的关系就不需要重复标注，减少冗余。

表 2 实体类型及相应说明

Table 2 Entity type and the corresponding description

实体类型	说明
人物	表示一个或多个人物的实体
设施	建筑、机场、铁路、桥等表示设施本身而非地点的实体
组织	表示一个或多个公司、机构、单位和组织等的实体
政治实体	一个或多个国家、城市、省（州）和行政区域等政治实体
地点	非政治实体的、表示地点和区域的实体
时间	表示日期、时刻（清晨、半夜或具体的几时几分）或时间跨度的实体
装备	表示一个或多个武器装备的实体

表 3 关系类型及相应说明

Table 3 Relation type and the corresponding description

关系类型	说明
隶属	人物或组织隶属于组织或政治实体、政治实体隶属于政治实体的关系
共事	两个人物隶属于同一个组织，则属于共事关系
归属	装备归属于组织或政治实体的关系
研制	组织或政治实体研制装备的关系
搭载	一个装备搭载人物或搭载另一个装备的关系
合作	人物、组织、政治实体之间的合作关系
敌对	人物、组织、政治实体之间的敌对关系
等价	重复出现的相同实体之间的等价关系

2 数据样本描述

数据集以 json 格式进行存储，每一条记录是一条军事新闻文本，json 数据的关键值其层级结构如图 2 所示。

其中：

(1) id 是新闻文本的序号。

(2) sentence 是经过清洗之后的新闻文本。

(3) tokens 是新闻文本经过分词后形成的 token 列表。

(4) tokens_count 是 token 个数。

(5) event_mention 是事件提及字典，非事件样本的 event_mention 为{}，event_mention 层中包含的 trigger 为事件触发器，trigger 层中包含的 text 为事件触发词文本，trigger 层中包含的 offset 为事件触发词在 token 列表中的偏移量；event_type 是标注的事件类型；arguments 是标注的事件类型对应的论元及其角色，arguments 层中包含的 role 是事件论元的角色，text 是事件论元文本，offset

是事件论元在 token 列表中的偏移量。

(6) entity_mention 是实体类型提及词典，entity_mention 层中包含的 entity_type 是实体类型，text 是实体文本，offset 是实体在 token 列表中的偏移量。

(7) relation_mention 是实体关系提及词典，relation_mention 层中包含的 head 是关联实体中的头实体，head 层中的 entity_type 是头实体类型，offset 是头实体在 token 列表中的偏移量；tail 是关联实体中的尾实体，tail 层中的 entity_type 是尾实体类型，offset 是尾实体在 token 列表中的偏移量；relation_type 是实体关系类型。

具体样本数据展示如图 3 所示。

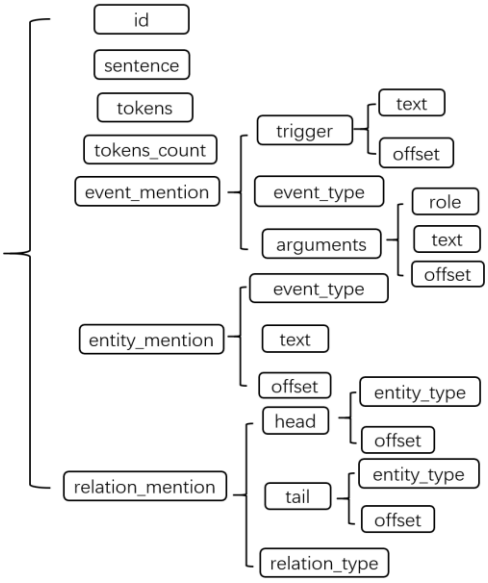

图 2 军事新闻文本数据集标注层次结构图

Figure 2 The labelled hierarchy diagram of military news text dataset


```
[{"id": "1", "sentence": "2012年底,国内军事论坛开始讨论一种称为055型的新型导弹驱逐舰,据鼎盛军事的爆料人士称,该驱逐舰的满载排水量达到了空前的12000吨。据悉,该舰将于2015年第二季度开工建造,目前该舰进入最后设计冲刺阶段。", "tokens": ["2012", "年底", ",", "国内", "军事论坛", "开始", "讨论", "一种", "称为", "055", "型", "的", "新型", "导弹驱逐舰", ",", "据", "鼎盛", "军事", "的", "爆料", "人士", "称", ",", "该", "驱逐舰", "的", "满载", "排水量", "达到", "了", "空前", "的", "12000", "吨", "。", "据悉", ",", "该舰", "将", "于", "2015", "年", "第二季度", "开工", "建造", ",", "目前", "该舰", "进入", "最后", "设计", "冲刺", "阶段", "。"], "tokens_count": 54, "event_mention": {}, "entity_mention": [{"entity_type": "TIME", "text": "2012年底", "offset": [0, 2]}, {"entity_type": "EQU", "text": "055型", "offset": [9, 11]}, {"entity_type": "EQU", "text": "新型导弹驱逐舰", "offset": [12, 14]}, {"entity_type": "ORG", "text": "鼎盛军事", "offset": [16, 18]}, {"entity_type": "PER", "text": "爆料人士", "offset": [19, 21]}, {"entity_type": "EQU", "text": "该驱逐舰", "offset": [23, 25]}, {"entity_type": "EQU", "text": "该舰", "offset": [37, 38]}, {"entity_type": "TIME", "text": "2015年第二季度", "offset": [40, 43]}], "relation_mention": [{"head": {"entity_type": "EQU", "offset": [9, 11]}, "tail": {"entity_type": "EQU", "offset": [12, 14]}, "relation_type": "Equal"}, {"head": {"entity_type": "PER", "offset": [19, 21]}, "tail": {"entity_type": "ORG", "offset": [16, 18]}, "relation_type": "Subordinate"}, {"head": {"entity_type": "EQU", "offset": [12, 14]}, "tail": {"entity_type": "EQU", "offset": [23, 25]}, "relation_type": "Equal"}, {"head": {"entity_type": "EQU", "offset": [23, 25]}, "tail": {"entity_type": "EQU", "offset": [37, 38]}, "relation_type": "Equal"}]},
```


图 3 军事新闻文本数据集样例

Figure 3 The sample of the military news text dataset

3 数据质量控制和评估

首先本研究所用到的数据来自新浪军事、凤凰军事和网易军事共三家公开的军事新闻网站，媒体新闻网站是信息自动化技术的有效语料来源，并且初步爬取了 40000 余条军事新闻，因此本研究

涉及的新闻文本内容以及新闻来源具有全面性、准确性和可靠性。

其次，为了进一步提高数据集的质量，在数据标注阶段将标注人员分为两组，每组 10 人，两组独立标注，对标注结果进行分析，对不一致的标注进行讨论，设计了：（1）事件类型准确率，（2）事件元素召回率，（3）命名实体召回率，（4）实体关系召回率 4 种质量指标（如表 4 所示）。整体标注结束后随机抽检一定数量的样本，若随机抽检指标均达 95%以上，则标注质量满足要求，否则继续改进标注流程至满足为止。经多轮迭代标注修改，本数据集最终的各指标结果如图 4 所示。

表 4 质量指标及相应说明

Table 4 Quality indicators and the corresponding description

指标	说明
事件类型准确率	（正确标注事件类型的样本个数/样本总数）×100%
事件元素召回率	（正确标注的事件元素个数/实际事件元素个数）×100%
命名实体召回率	（正确标注的命名实体个数/实际命名实体个数）×100%
实体关系召回率	（正确标注的实体关系个数/实际实体关系个数）×100%

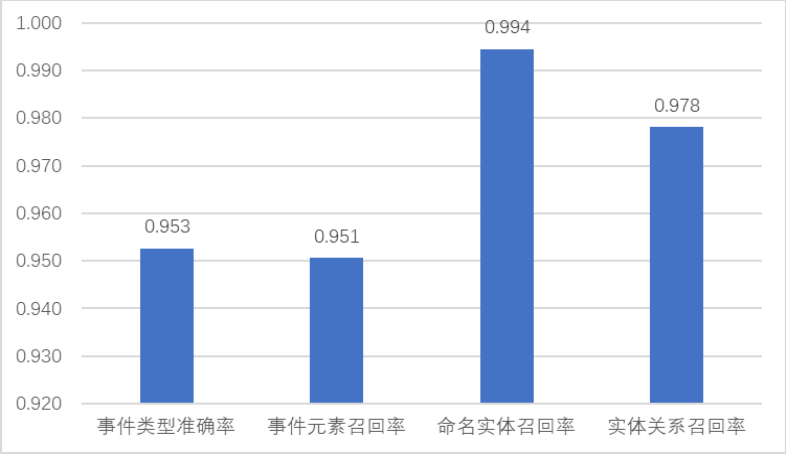

图 4 质量指标结果展示

Figure 4 Display of quality indicator results

最后，对完成数据标注之后的数据集设计了：（1）事件类型分布统计本总数，（2）事件元素完整度统计，（3）命名实体类型分布统计，（4）实体关系类型分布统计 4 种统计指标，从而全面了解数据集中事件分布情况，保证数据的有效性。统计指标说明如表 5 所示。

表 5 统计指标及相应说明

Table 5 Statistical indicators and the corresponding description

指标	说明
事件类型分布统计本总数	统计 8 类事件类型的样本个数
事件元素完整度统计	除了非事件的 7 类事件类型的事件元素完整度
命名实体类型分布统计	各类型命名实体的个数和包含该类型命名实体（1 个及以上）的样本个数
实体关系类型分布统计	各类型实体关系的个数和包含该类型实体关系的样本个数

最终 13000 条可用于军事领域的军事新闻文本数据集事件类型分布如表 6 所示，命名实体类型分布如表 7 所示，实体关系类型分布如表 8 所示。

表 6 事件类型分布及事件元素完整度

Table 6 Event type distribution and event element completeness

事件类型	样本个数	事件元素完整度
非事件（Non-event）	6003	——
试验（Experiment）	907	0.772
演习（Manoeuvre）	4299	0.828
展示（Exhibit）	975	0.839
部署（Deploy）	287	0.883
意外事故（Accident）	296	0.881
保障（Indemnity）	75	0.917
支援（Support）	158	0.947

表 7 命名实体类型分布

Table 7 Entity type distribution and event element completeness

实体类型	实体个数	包含该类实体的样本个数
时间（TIME）	16410	10510
政治实体（GPE）	10480	6137
装备（EQU）	28792	9549
人物（PER）	10213	6115
组织（ORG）	21280	10469
地点（LOC）	9619	6444
设施（FAC）	714	498

表 8 实体关系类型分布

Table 8 Entity relation type distribution and event element completeness

关系类型	关系个数	包含该类关系的样本个数
归属（Belong）	12705	6783
搭载（Carry）	3008	1955
隶属（Subordinate）	11744	6607
等价（Equal）	8903	5828
敌对（Hostility）	428	377
研制（Develop）	1592	1172
合作（Cooperation）	1701	1099
共事（Fellow）	171	130

4 数据价值

事件抽取技术应用于军事情报领域，可以大大提升情报获取的效率，不仅能动态、实时地进行情报库的扩展，还能构建情报间的联系、揭示情报间的时序性，根据时间发展描摹重大事件的进程，

为后续的军事决策提供坚实的基础[14]。在军事领域数字化进程不断推进的背景之下，如何有效地提炼、管理和利用军事信息，对于提高军事资源的利用效率和军事知识的应用研究具有重要意义[15]。对军事新闻进行事件抽取，可以帮助决策者从海量新闻中有效地提取军事事件类型和事件元素，快速地挖掘出有价值的新闻信息，更好地了解各方军队的最新动向。由于军事领域的特殊性，目前国内外尚缺乏权威高质且对外开放的用于事件抽取的军事新闻数据集资源，军事新闻文本数据集资源非常匮乏，因此构建相应的数据集尤为重要，具有很大的现实应用价值[12]。

本文构建的军事新闻文本数据集包含 13000 条高质量新闻文本，聚焦军事事件建模，能够有效地缓解军事新闻数据集匮乏这一问题，包含了试验、演习、部署、保障、支援、意外事故、展示、非事件等 8 种事件类型，以及人物、设施、组织和装备等 7 种实体类型和隶属、共事、合作和敌对等 8 种实体关系，是一个标注种类齐全、质量高的军事新闻文本数据集，有助于推动军事领域事件抽取的研究，推动该领域的发展。

数据作者分工职责

黄宏斌（1975—），男，江苏省如皋市人，博士，研究员，研究方向为信息管理与智能决策、军事信息物理系统（MCPS）、大数据分析。主要承担工作：本数据集构造方法的指导和论文写作指导，数据质量控制与综合管理。

孙皎（1999—），女，安徽省合肥市人，硕士研究生，研究方向为信息管理与智能决策。主要承担工作：数据资料的搜集，数据的预处理和整合，数据挑选、数据校对，论文撰写。

韦晖（1996—），男，广西壮族自治区南宁市人，硕士研究生，研究方向为信息管理与智能决策。主要承担工作：数据资料的搜集与构建，数据挑选、数据校对，论文修改。

肖开明（1990—），男，云南省腾冲市人，博士，讲师，研究方向为大数据技术、信息管理与智能决策。主要承担工作：数据集选择与采集指导，数据集构造方法研究和论文写作指导，数据集质量控制。

王懋（1989—），男，湖南省常德市人，博士，讲师，研究方向为数据集成。主要承担工作：数据集选择与构建指导，论文撰写指导。

李璇（1982—），女，湖南省长沙市人，博士，副研究员，研究方向为数据工程。主要承担工作：数据集选择与构建指导，论文撰写指导。

参考文献

[1] 秦彦霞, 张民, 郑德权. 神经网络事件抽取技术综述[J]. 智能计算机与应用, 2018, 8(3): 1–5, 10. DOI:10.3969/j.issn.2095-2163.2018.03.002. [QIN Y X, ZHANG M, ZHENG D Q. A survey on neural network-based methods for event extraction[J]. Intelligent Computer and Applications, 2018, 8(3): 1–5, 10. DOI:10.3969/j.issn.2095-2163.2018.03.002.]

[2] 高李政, 周刚, 罗军勇, 等. 元事件抽取研究综述[J]. 计算机科学, 2019, 46(8): 9–15. DOI: 10.11896/j.issn.1002-137X.2019.08.002. [GAO L Z, ZHOU G, LUO J Y, et al. Survey on meta-event extraction[J]. Computer Science, 2019, 46(8): 9–15. DOI: 10.11896/j.issn.1002-137X.2019.08.002.]

[3] 高鸿博. 面向百科数据的军事事件抽取[D].大连理工大学, 2021. [GAO H B. Military Event Extraction for Encyclopedia Data[D]. Dalian University of Technology, 2021.]

[4] WALKER C, STRASSEL S, MEDERO J, et al. ACE 2005 multilingual training corpus LDC2006T06[M/OL]. Philadelphia: Linguistic Data Consortium, 2006. https://catalog.ldc.upenn.edu/LDC2006T06.]

[5] WANG X Z, WANG Z Q, HAN X, et al. MAVEN: a massive general domain event detection dataset[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 1652–1671. DOI: 10.18653/v1/2020.emnlp-main.129.

[6] ZHU H Q, XIAO K M, OU L Z, et al. Attention-based graph convolution networks for event detection[C]//7th International Conference on Big Data and Information Analytics (BigDIA). Chongqing, China. IEEE, 2021: 185–190. DOI: 10.1109/BigDIA53151.2021.9619647.

[7] 潘丽敏, 李筱雅, 罗森林, 等. 融合多级语义特征的双通道 GAN 事件检测方法[J]. 北京理工大学学报, 2021, 41(3): 295–302. DOI: 10.15918/j.tbit1001-0645.2019.177. [PAN L M, LI X Y, LUO S L, et al. Double-channel GAN with multi-level semantic correlation for event detection[J]. Transactions of Beijing Institute of Technology, 2021, 41(3): 295–302. DOI: 10.15918/j.tbit1001-0645.2019.177.]

[8] 田梓函, 李欣. 基于 BERT-CRF 模型的中文事件检测方法研究[J]. 计算机工程与应用, 2021, 57(11): 135–139. DOI: 10.3778/j.issn.1002-8331.2006-0065. [TIAN Z H, LI X. Research on Chinese event detection method based on BERT-CRF model[J]. Computer Engineering and Applications, 2021, 57(11): 135–139. DOI: 10.3778/j.issn.1002-8331.2006-0065.]

[9] 潘璋, 黄德根. 事件要素注意力与编码层融合的触发词抽取研究[J]. 小型微型计算机系统, 2021, 42(4): 673–677. DOI: 10.3969/j.issn.1000-1220.2021.04.001. [PAN Z, HUANG D G. Research on trigger word extraction based on the fusion of event argument attention and encoder layer[J]. Journal of Chinese Computer Systems, 2021, 42(4): 673–677. DOI: 10.3969/j.issn.1000-1220.2021.04.001.]

[10] DODDINGTON G, MITCHELL A, PRZYBOCKI M A, et al. The automatic content extraction (ACE) program - tasks, data, and evaluation[C]//Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004. European Language Resources Association, 2004: 837–840.

[11] 游飞. 基于深度学习的军事事件抽取研究[D]. 中国电子科技集团公司电子科学研究院, 2018. [YOU F. Research on Military Event Extraction Based on Deep Learning[D]. China Academic of Electronics and Information Technology, 2018.]

[12] 战立莹. 军事想定事件抽取关键技术研究[D]. 国防科技大学, 2019. [ZHAN L Y. Research on the Key Techologies of Military Scenario Event Extraction[D]. National University of Defense Technology, 2019.]

[13] CHE W X, LI Z H, LIU T. LTP: a Chinese language technology platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Beijing, China. New York: ACM, 2010: 13–16. DOI: 10.5555/1944284.1944288.

[14] 柳亦婷. 面向特定领域的事件抽取研究[D]. 苏州大学, 2018. [LIU Y T. Research on domain-specific event extraction[D]. Soochow University, 2018.]

[15] 薛坤. 面向军事领域的知识图谱构建与应用研究[D].大连理工大学, 2020. [XUE K. Research on the Construction and Application of Knowledge Graph in the Military Field[D]. Dalian University of Technology, 2020.]

论文引用格式

黄宏斌，孙皎，韦晖，等. 基于开源军事新闻的领域事件数据集[J/OL]. 中国科学数据，2023，8(1). (2023-03-10). DOI: 10.11922/11-6035.csd.2022.0072.zh.

数据引用格式

黄宏斌，孙皎，韦晖，等. 基于开源军事新闻的领域事件数据集[DS/OL]. Science Data Bank，2022. (2022-09-06). DOI: 10.57760/sciencedb.j00001.00486.

A dataset of domain events based on open-source military news

HUANG Hongbin[1*], SUN Jiao[1], WEI Hui[1], XIAO Kaiming[1], WANG Mao[1], LI Xuan[1]

1. Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410073, P. R. China

*Email: hbhuang@nudt.edu.cn

Abstract: The text dataset of the military field is the basis for event extraction in the military field, and the datasets of high quality can effectively promote the study of event extraction in this field. However, the event extraction dataset commonly used in the real world (such as ACE2005, etc.) is oriented to the general field, and the text corpus resources on military events are scarce. Therefore, we collected a large amount of military news content from public military news websites. Firstly, on the basis of text content analysis, we first established an event model of military news including event types, entity types and entity relationship types. Secondly, we manually labeled the text data according to the event model, which was iteratively verified and corrected simultaneously. Finally, we obtained dataset of 13,000 high-quality military news events with a full variety of labels. We make this military news event dataset publicly available in this paper.

Keywords: natural language processing; event extraction; military news; dataset

Dataset Profile

Title	A dataset of domain events based on open-source military news
Data corresponding author	HUANG Hongbin (hbhuang@nudt.edu.cn)
Data author(s)	HUANG Hongbin, SUN Jiao, WEI Hui, XIAO Kaiming, WANG Mao, LI Xuan
Time range	2010–2022
Data volume	25.6 MB
Data format	*.json
Data service system	http://dx.doi.org/10.57760/sciencedb.j00001.00486
Dataset composition	The dataset contains 13,000 military news, stored in *.json format, with a data volume of 25.6 MB.