ICS
CCS

DB65

新 疆 维 吾 尔 自 治 区 地 方 标 准

DB XX/T XXXX—XXXX

多语言文本信息抽取语料库标注规则及建设要求

Annotation Rules and Construction Requirements for Multilingual Text Information Extraction Corpora

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

2023－XX－XX 发布　　2023－XX－XX 实施

新疆维吾尔自治区市场监督管理局　　发 布

目　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国科学院新疆理化技术研究所提出。

本文件由新疆维吾尔自治区工业和信息化厅归口并组织实施。

本文件起草单位：中国科学院新疆理化技术研究所。

本文件主要起草人：XX。

本文件实施应用中的疑问，请咨询中国科学院新疆理化技术研究所。

对本文件的修改意见及建议，请反馈至中国科学院新疆理化技术研究所（乌鲁木齐市新市区科学二街181号）、新疆维吾尔自治区市场监督管理局（乌鲁木齐市新华南路167号）。

中国科学院新疆理化技术研究所 联系电话：0991-3837795；传真：0991-3838957；邮编：830046。

新疆维吾尔自治区市场监督管理局 联系电话：0991-2818750；传真：0991-2311250；邮编：830004。

多语言文本信息抽取语料库标注规则及建设要求

1 范围

本文件规定了多语言文本信息抽取语料库建设所需的术语和定义、标注信息元数据、标注流程。

本文件适用于多语言文本信息抽取语料库的采集、标注与建设，信息抽取任务包括命名实体识别、关系抽取、事件抽取、话题检测及跟踪，多语言文本信息抽取语料库标注及建设的同类语料库标注及建设参照执行。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867-2022 信息技术 人工智能 术语

GB/T 42131-2022 人工智能 知识图谱技术框架

GM/T 0125.1-2022 JSON Web 密码应用语法规范 第1部分：算法标识

3 术语和定义

下列术语和定义适用于本文件。

3.1

语料 corpus

语言材料或资料。

[来源：GB/T 40035-2021，3.2]。

3.2

语料库 corpora

集中起来供分析用的语料集合。

[来源：GB/T 15237.1—2000，3.6.9, 有修改]

3.3

多语言 multilingual

指信息抽取语料库包含多种语言的语料，且本标准可同时适用于多种语言。

3.4

多语言语料库 multilingual corpora

由两种或两种以上语言语料组成的语料库。

3.5

信息抽取 information extraction

把文本里包含的信息进行结构化处理，形成类似表格的组织形式。

3.6

实体 entity
所关注的具体或抽象事物，包括事物之间的关联。
[来源：GB/T 41867-2022，3.3.7]

3.7

命名实体 named entity
具有特定或唯一含义的指称名称的实体。
注：指称名称包括人员、位置、组织的特定名称，以及基于域或应用的其他专有名称。
[来源：GB/T 41867-2022，3.3.7]

3.8

命名实体识别 named entity recognition
识别并标注文本中单词序列的实体的指称名称及其类别的任务。
[来源：GB/T 41867-2022，3.3.7，有修改]

3.9

关系 relation
实体、实体类型、实体组合或实体类型组合间的联系。
注：关系可描述实体类型和实体类型、实体类型和实体、实体和实体之间的关联方式。
[来源：GB/T 42131-2022，3.11]

3.10

关系抽取 relation extraction
识别文本中提到的实体之间关系的任务。
[来源：GB/T 41867-2022，3.3.4]

3.11

事件 event
发生在某个特定时间点或时间段、某个特定地域范围内，由一个或者多个角色参与的一个或者多个动作组成的事情或者状态的改变。
注：可表示为具有时间属性的实体和关系的组合。
[来源：GB/T 42131-2022，3.12]

3.12

事件抽取 event extraction
从语料（3.1）中抽取特定类型的事件（3.11）以及相关实体（3.6）信息，并形成结构化数据。

3.13

话题 topic
日常生活中被关注和讨论的一类事件（3.11）内容的一个概括。

3.14

话题检测与跟踪 topic detection and tracking
对新闻等媒体信息流进行新话题的自动识别和已知话题（3.13）的持续跟踪。

3.15

JSON Javascript Object Notation
Javascript对象标记，一种轻量级、基于文本的、语言独立的数据交换格式。
[来源：GM/T 0125.1-2022，3.1]

4 标注信息元数据

4.1 基本信息

注：基本信息元数据用于存储信息抽取语料库的9种基本信息，包括编号、领域、创建人、创建时间、描述、状态、版本、规模、来源，并使用JSON格式文件进行单独存储。

4.1.1 编号

中文名称：编号
英文名称：id
定　　义：语料库的唯一标识符
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例：“202105060001”

4.1.2 领域

中文名称：领域
英文名称：domain
定　　义：语料库内容所属的专业领域
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例：“政治”

4.1.3 创建人

中文名称：创建人
英文名称：creator
定　　义：语料库创建机构或个人名称
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例：“中国科学院新疆理化技术研究所”

4.1.4 创建时间

中文名称：创建时间
英文名称：create_time
定　　义：语料库的创建时间
数据类型：字符串
值　　域：用阿拉伯数字将年、月、日标全，具体格式应为YYYY-MM-DD HH:mm:ss
是否必填：是
取值示例：“2021-05-06 11:12:38”

4.1.5 描述

中文名称：描述

英文名称：description
定　　义：语料库内容的简要描述
数据类型：字符串
值　　域：无要求
是否必填：否
取值示例：“新闻文本”

4.1.6　状态

中文名称：状态
英文名称：status
定　　义：语料库文件的标注状态
数据类型：整数
值　　域：“标注中”、“标注完成”、“标注失败”
是否必填：是
取值示例：“标注完成”

4.1.7　版本

中文名称：版本
英文名称：version
定　　义：语料库的版本信息
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例：“V1.0”

4.1.8　规模

中文名称：规模
英文名称：size
定　　义：语料库中语料的总数量
数据类型：整数
值　　域：无要求
是否必填：是
取值示例：1024

4.1.9　来源

中文名称：来源
英文名称：origin
定　　义：语料库数据的来源信息
数据类型：字符串
值　　域：无要求
是否必填：否
取值示例：“人民网”

4.2　命名实体

4.2.1　索引

中文名称：索引
英文名称：index
定　　义：当前语料在语料库中的索引值
数据类型：整数
值　　域：0至语料规模减1
是否必填：是
取值示例：0

4.2.2 文本

中文名称：文本
英文名称：text
定　　义：当前语料的文本内容
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例：“习近平在新疆乌鲁木齐考察调研”

4.2.3 语言

中文名称：语言
英文名称：language
定　　义：当前语料的语言种类
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例：“中文”

4.2.4 实体列表

中文名称：实体列表
英文名称：entity_list
定　　义：当前语料标注出来的所有实体结构体组成的列表
数据类型：列表
值　　域：无要求
是否必填：是
取值示例：[{"entity_index": 0, "entity_text": "新疆", "entity_type": "地名", "entity_start_position":4, "entity_end_position":5}]

4.2.5 实体结构体

中文名称：实体结构体
英文名称：entity_structure
定　　义：当前语料标注出来的实体
数据类型：结构体
值　　域：无要求
是否必填：否
取值示例：{"entity_index": 0, "entity_text": "新疆", "entity_type": "地名", "entity_start_position":4, "entity_end_position":5}

注　　解：实体结构体内包含实体索引、实体文本、实体类型、实体开始位置、实体结束位置5个字段，并分别采用4.2.6-4.2.10的定义

4.2.6　实体索引

中文名称：实体索引
英文名称：entity_index
定　　义：当前实体在当前语料所有实体中的索引值
数据类型：整数
值　　域：0至当前语料实体总数量减1
是否必填：是
取值示例：0

4.2.7　实体文本

中文名称：实体文本
英文名称：entity_text
定　　义：当前实体的文本内容
数据类型：字符串
值　　域：当前语料包含的字符串
是否必填：是
取值示例：“新疆”
注　　解：维吾尔语等语言存在后缀，应去除不标注进实体文本

4.2.8　实体类型

中文名称：实体类型
英文名称：entity_type
定　　义：当前实体所属的实体类别
数据类型：字符串
值　　域：无
是否必填：是
取值示例：“地名”

4.2.9　实体开始位置

中文名称：实体开始位置
英文名称：entity_start_position
定　　义：当前实体文本在语料中的开始位置
数据类型：整数
值　　域：0至语料文本总字符数减1
是否必填：是
取值示例：0
注　　解：实体位置标注采用字符级的位置；语言书写顺序包括从左往右（如中文）和从右往左（如维吾尔语）两种，实体位置的计算顺序与该语料语言的书写顺序保持一致

4.2.10　实体结束位置

中文名称：实体结束位置
英文名称：entity_end_position

定　　义：当前实体文本在语料中的结束位置
数据类型：整数
值　　域：0至语料文本总字符数减1
是否必填：是
取值示例：1
注　　解：同4.2.9注解

4.3 关系

注 1：关系元数据也包括语料索引、文本、语言、实体列表、实体结构体、实体索引、实体文本、实体开始位置、实体结束位置、实体类型，并使用命名实体元数据 4.2 中对应的定义方法。

注 2：实体结构体中的实体类型不是必填项。

4.3.1 关系列表

中文名称：关系列表
英文名称：relation_list
定　　义：当前语料标注出来的所有关系结构体组成的列表
数据类型：列表
值　　域：无要求
是否必填：是
取值示例：[{"relation_index": 1, "relation_type": "位于", "head_entity_index":1, "tail_entity_index":0}]

4.3.2 关系结构体

中文名称：关系结构体
英文名称：relation_structure
定　　义：当前语料标注出来的关系三元组
数据类型：结构体
值　　域：无要求
是否必填：否
取值示例：{"relation_index": 1, "relation_type": "位于", "head_entity_index":1, "tail_entity_index":0}
注　　解：关系结构体内包含关系索引、关系类型、头实体索引、尾实体索引4个字段，并分别采用4.3.3-4.3.6的定义

4.3.3 关系索引

中文名称：关系索引
英文名称：relation_index
定　　义：当前关系在当前语料所有关系中的索引值
数据类型：整数
值　　域：0至当前语料关系总数量减1
是否必填：是
取值示例：0

4.3.4 关系类别

中文名称：关系类别
英文名称：relation_type

定　　义：关系三元组的类型
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例：“位于”

4.3.5 头实体索引

中文名称：头实体索引
英文名称：head_entity_index
定　　义：关系三元组中的头实体
数据类型：整数
值　　域：0至当前语料实体总数量减1
是否必填：是
取值示例：1

4.3.6 尾实体索引

中文名称：尾实体索引
英文名称：tail_entity_index
定　　义：关系三元组中的尾实体
数据类型：整数
值　　域：0至当前语料实体总数量减1
是否必填：是
取值示例：0

4.4 事件

注：事件元数据也包括索引、文本、语言，并使用命名实体元数据 4.2 中对应的定义方法。

4.4.1 事件列表

中文名称：事件列表
英文名称：event_list
定　　义：当前语料标注出来的所有事件结构体组成的列表
数据类型：列表
值　　域：无要求
是否必填：是
取值示例：[{"event_index":0, "event_type":"考察", "event_trigger": {"entity_text": "考察", "entity_start_position":10, "entity_end_position":11}, "event_argument_list": [{"argument_role": "考察人", "entity_text":"习近平", "entity_start_position":0, "entity_end_position":2}, {"argument_role": "考察地点", "entity_text":"新疆乌鲁木齐", "entity_start_position":4, "entity_end_position":9}]}]

4.4.2 事件结构体

中文名称：事件结构体
英文名称：event_structure
定　　义：当前语料标注出来的事件结构体
数据类型：结构体
值　　域：无要求

是否必填：否

注　　解：事件结构体内包含事件索引、事件文本、事件触发词、事件要素列表4个字段，并分别采用4.4.3-4.4.6的定义

取值示例：{"event_index":0, "event_type":"考察", "event_trigger": {"entity_text": "考察", "entity_start_position":10, "entity_end_position":11}, "event_argument_list": [{"argument_role": "考察人", "entity_text":"习近平", "entity_start_position":0, "entity_end_position":2}, {"argument_role": "考察地点", "entity_text":"新疆乌鲁木齐", "entity_start_position":4, "entity_end_position":9}]}

4.4.3　事件索引

中文名称：事件索引

英文名称：event_index

定　　义：事件结构体在当前语料的所有事件结构体中的索引值

数据类型：整数

值　　域：0至当前语料事件结构体数量减1

是否必填：是

取值示例：0

4.4.4　事件类型

中文名称：事件类型

英文名称：event_type

定　　义：事件结构体的事件类型

数据类型：字符串

值　　域：无要求

是否必填：是

取值示例：“考察”

4.4.5　事件触发词

中文名称：事件触发词

英文名称：event_trigger

定　　义：事件结构体的事件触发词

数据类型：结构体

值　　域：无要求

是否必填：是

注　　解：事件触发词结构体内包含触发词文本、触发词开始位置、触发词结束位置三个字段，并分别采用4.2.7、4.2.8、4.2.9的定义

取值示例：{"entity_text": "考察", "entity_start_position":10, "entity_end_position":11}

4.4.6　事件要素列表

中文名称：事件要素列表

英文名称：event_argument_list

定　　义：当前事件包含的所有事件要素结构体组成的列表

数据类型：列表

值　　域：无要求

是否必填：是

取值示例：[{"argument_role": "考察人", "entity_text":"习近平", "entity_start_position":0, "entity_end_position":2}, {"argument_role": "考察地点", "entity_text":"新疆乌鲁木齐", "entity_start_position":4, "entity_end_position":9}]

4.4.7 事件要素

中文名称：事件要素
英文名称：event_argument
定　　义：当前事件结构体包含的事件要素
数据类型：结构体
值　　域：无要求
是否必填：否
注　　解：事件要素结构体包含要素角色、要素文本、要素文本开始位置、要素文本结束位置四个字段，其中要素文本、要素开始位置、要素结束位置分别采用4.2.7、4.2.8、4.2.9的定义
取值示例：{"argument_role": "考察人", "entity_text":"习近平", "entity_start_position":0, "entity_end_position":2}

4.4.8 要素角色

中文名称：要素角色
英文名称：argument_role
定　　义：当前事件要素在事件结构体中的角色类型，如时间、地点等
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例：“考察人”

4.5 话题

注1：话题元数据也包括语料索引、语料语言，并使用命名实体元数据4.2中对应的定义方法。

4.5.1 文本列表

中文名称：文本列表
英文名称：text_list
定　　义：由多条文本组成的文本列表
数据类型：列表
值　　域：无要求
是否必填：是
取值示例：["习近平在新疆乌鲁木齐考察调研", "习近平总书记在新疆乌鲁木齐市考察了新疆大学", "阿根廷获得卡塔尔世界杯冠军", "阿根廷队与法国队在常规时间和加时赛战成3比3平"]

4.5.2 话题列表

中文名称：话题列表
英文名称：topic_list
定　　义：当前语料标注出来的所有话题组成的列表
数据类型：列表
值　　域：无要求
是否必填：是

取值示例：[{"topic_index":0, "topic_type":"政治", "topic_start_position":0, "topic_end_position":1}, {"topic_index":1, "topic_type":"足球", " topic_start_position":2, " topic_end_position":3}]

4.5.3 话题结构体

中文名称：话题结构体
英文名称：topic_structure
定　　义：当前语料包含的话题
数据类型：结构体
值　　域：无要求
是否必填：否
注　　解：话题结构体内包含话题索引、话题类型、话题开始位置、话题结束位置四个字段，并分别采用4.5.4、4.5.5、4.5.6的定义
取值示例：{"topic_index":0, "topic_type":"政治", "topic_start_position":0, "topic_end_position":1}, {"topic_index":1, "topic_type":"足球", " topic_start_position":2, " topic_end_position":3}

4.5.4 话题索引

中文名称：话题索引
英文名称：topic_index
定　　义：当前话题在当前语料所有话题中的索引值
数据类型：整数
值　　域：0至语料总话题数减1
是否必填：必选
取值示例：0

4.5.5 话题类型

中文名称：话题类型
英文名称：topic_type
定　　义：当前话题所属的话题类别
数据类型：字符串
值　　域：无要求
是否必填：必选
取值示例：“政治”

4.5.6 话题开始位置

中文名称：话题开始位置
英文名称：topic_start_position
定　　义：当前话题在语料文本列表中以句子为单位的开始位置
数据类型：整数
值　　域：0至语料文本列表中句子总数量减1
是否必填：必选
取值示例：0

4.5.7 话题结束位置

中文名称：话题结束位置
英文名称：topic_end_position

定　　义：当前话题在语料文本列表中以句子为单位的结束位置
数据类型：整数
值　　域：0至语料文本列表中句子总数量减1
是否必填：必选
取值示例：1

5　建设流程

数据建设工作流程见图1。

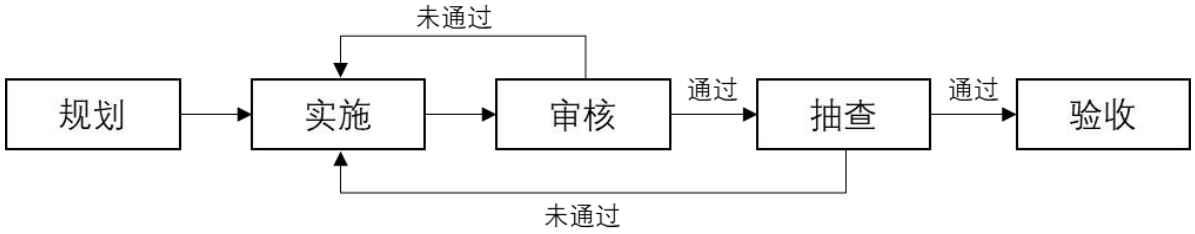

图 1 数据建设工作流程图

5.1　规划

5.1.1　明确需求

标注工作开始之前，数据标注项目的负责人应及时获取数据需求方的数据标注说明书，应包括以下内容：

1) 明确标注内容和标注规则；
2) 明确标注时间节点
3) 明确标注验收规则；

5.1.2　明确计划

数据标注项目负责人应根据标注需求制定标注计划，包括进度计划、人员计划、资金计划、工具计划、质量控制计划、验收计划等。

5.1.3　专项培训

按照标注计划和标注规则，对数据标注人员进行有针对性的培训，确保标注质量。

5.2　实施

5.2.1　数据准备

收集和准备需要标注的原始数据，确保数据的完整性、准确性和可用性，并对数据进行清洗、去噪和预处理等操作。

5.2.2　任务创建

负责人利用标注工具创建数据标注项目相关内容。

5.2.3　任务分发

根据标注需求和数据量，负责人将数据标注任务分派给数据标注人员。

5.2.4 任务实施

数据标注人员使用相应数据标注工具完成指派的数据标注任务。

5.3 审核

5.3.1 制定审核标准

确定明确的审核标准和指标，确保标注结果与预期结果一致。

5.3.2 数据标注验证

数据标注审查人员对标注好的数据进行审核，按标注要求比对标注结果，以确保准确性和一致性。

5.3.3 任务数据回收

对标注不合格数据进行回收，并重新分派进行标注。

5.3.4 问题记录与反馈

在审核过程中记录发现的问题和错误，并与标注团队进行问题反馈与沟通，确保标注团队理解存在的问题和改进要求。

5.4 抽查

5.4.1 随机抽样验证

从标注好的数据中随机选择一部分样本进行检查，并按标注要求比对标注结果，以确保准确性和一致性。

5.4.2 任务数据回收

对标注不合格数据进行回收，并重新分派进行标注。

5.4.3 问题记录与反馈

在审核过程中记录发现的问题和错误，并与标注团队进行问题反馈与沟通，确保标注团队理解存在的问题和改进要求。

5.5 验收

5.5.1 确定验收标准

需求方需要明确数据标注的验收标准和指标，例如数据的格式、质量、准确性等要求，以及与标注任务相关的特定要求。

5.5.2 数据整理与交付

对标注数据进行整理并转换为JSON格式，确保数据的结构化和可访问性，同时提供详细的交付文件和文档，例如数据标注说明、数据格式说明、标签定义等。

5.5.3 验收报告和文档

撰写验收报告，记录验收过程、问题、改进措施和结果，报告应包括标注数据的准确性、一致性、完整性等评估结果。

附 录

数据标注结果导出与交付可采用JSON文件格式，采用UTF-8编码方式，导出格式示例如下。

1. 数据库基本信息 JSON 格式示例

该 JSON 文件用于存储多语言信息抽取语料库的基本信息，文件名可为 corpora_info.json。注：每个信息抽取任务标注产生的语料库应包含对应语料库基本信息的 JSON 文件。

实例：

```
{
    "id": "202105060001",
    "domain": "政治",
    "creator": "中国科学院新疆理化技术研究所",
    "create_time": "2021-05-06 11:12:38",
    "description": "新闻文本",
    "status": "标注完成",
    "version": "V1.0",
    "size": 1024,
    "origin": "人民网"
}
```

2. 命名实体标注 JSON 格式示例

该 JSON 文件用于存储命名实体识别标注任务的语料库标注结果，文件名可为 corpora_ner.json。

中文实例：

```
{
    "index": 0,
    "text": "习近平在新疆乌鲁木齐考察调研",
    "language": "中文",
    "entity_list": [
        {
            "entity_index": 0,
            "entity_text": "新疆",
            "entity_type": "地名",
            "entity_start_position":4,
            "entity_end_position":5
        }
```

```
    ]
}
…
维吾尔语实例：
{
    "index": 0,
    "text": "باش شۇجى شى جىنپىڭ شىنجاڭ ئۇيغۇر ئاپتونوم رايونىنىڭ ئۈرۈمچى شەھىردە خىزمەت تەكشۈردى.",
    "language": "维语",
    "entity_list": [
        {
            "entity_index": 0,
            "entity_text": "شى جىنپىڭ ",
            "entity_type": "地名",
            "entity_start_position":9,
            "entity_end_position":17
        }
        …
    ]
}
…
```

3. 关系抽取标注 JSON 格式示例

该 JSON 文件用于存储关系抽取标注任务的语料库标注结果，文件名可为 corpora_relation.json。

```
中文实例：
{
    "index": 0,
    "text": "习近平在新疆乌鲁木齐考察调研",
    "language": "中文",
    "entity_list": [
        {
            "entity_index": 0,
            "entity_text": "新疆",
            "entity_start_position":4,
```

```
            "entity_end_position":5
        },
        {
            "entity_index": 1,
            "entity_text": "乌鲁木齐",
            "entity_start_position":6,
            "entity_end_position":9
        }
    ],
    "relation_list": [
        {
            "relation_index": 1,
            "relation_type": "位于",
            "head_entity_index":1,
            "tail_entity_index":0
        }
    ]
}
…
```


4. **事件抽取标注 JSON 格式示例**

该 JSON 文件用于存储事件抽取标注任务的语料库标注结果，文件名可为 corpora_event.json。


```
中文实例：
{
    "index":0,
    "text":"习近平在新疆乌鲁木齐考察调研",
    "language":"中文",
    "event_list":[{
        "event_index":0,
        "event_type":"考察",
        "event_trigger":
            {
                "entity_text": "考察",
                "entity_start_position":10,
```

```
                "entity_end_position":11
            },
        "argument_list": [
            {
                "argument_role": "考察人",
                "entity_text":"习近平",
                "entity_start_position":0,
                "entity_end_position":2

            },
            {
                "argument_role": "考察地点",
                "entity_text":"新疆乌鲁木齐",
                "entity_start_position":4,
                "entity_end_position":9

            }
        ]
    }]
}
…
```


5. 话题检测与跟踪标注 JSON 格式示例

该 JSON 文件用于存储话题检测与跟踪任务的语料库标注结果，文件名可为 corpora_topic.json。

中文实例：

```
{
    "index":0,
    "text_list": ["博文：习近平在新疆乌鲁木齐考察调研", "博文：习近平总书记在新疆乌鲁木齐市考察了新疆大学","博文：阿根廷获得卡塔尔世界杯冠军","博文：阿根廷队与法国队在常规时间和加时赛战成3比3平"],
    "language":"中文",
    "topic_list":[
        {
            "topic_index":0,
```

```
            "topic_type":"政治",
            "topic_start_position":0,
            "topic_end_position":1
        },
        {
            "topic_index":1,
            "topic_type":"足球",
            "topic_start_position":2,
            "topic_end_position":3
        }
    ]
}
…
```