ICS 35.240
CCS L60

T/CAPT

团　　体　　标　　准

T/CAPT 003—2021

中文新闻信息结构化标注规范

2021－10－19 发布　　2021－10－20 实施

中国新闻技术工作者联合会　　发 布

目　　次

前言 II
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　标注体系分类 1
5　实体类信息 1
5.1　实体类信息类型 1
5.2　实体类信息详情 2
6　业务类信息 3
6.1　业务类信息类型 3
6.2　业务类信息详情 3
7　多媒体元素类信息 6
7.1　多媒体元素类信息类型 6
7.2　多媒体元素类信息详情 6
附录 A（资料性）　传感器新闻信息 8

前　言

本文件参照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。

本文件由中国新闻技术工作者联合会新闻信息标准化分会秘书处和新华通讯社通信技术局联合提出。

本文件由中国新闻技术工作者联合会归口。

本文件起草单位：新华通讯社通信技术局、北京语言大学信息科学学院、中国人民大学新闻学院、北京星尘纪元智能科技有限公司、新华社媒体融合生产技术与系统国家重点实验室。

本文件主要起草人：王熠、饶高琦、唐铮、秦玉芳、徐铭锴、钱青青、郜沁清、杨溟、付蓉、刘一闻、王立欢。

中文新闻信息结构化标注规范

1 范围

本文件规定了中文新闻信息结构化标注的要求、标注规则和方法。

本文件适用于中文新闻领域信息内容的标注，服务于新闻信息资产的分析挖掘、知识发现和再利用，为多维度检索、组成特定专题、关系图谱等积累数据基础，为新闻信息内容的人工标注、半自动化及自动化标注应用提供指导和参考依据。

本文件的使用对象包括报刊、广播、电视、通讯社、新闻网站等新闻内容提供商及媒体应用与研究机构。

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 20092 中文新闻信息置标语言

GB/T 20093 中文新闻信息分类与代码

3 术语和定义

下列术语和定义适用于本文件。

3.1

策划类型

新闻内容与当前报道的新闻专题的关系。

注：与新闻专题报道直接相关的称为程序稿，与新闻专题报道进行周边报道和反馈的称为配合稿。

4 标注体系分类

中文新闻信息结构化标注通过分析总结新闻内容数据，按照标注体系划分为三个类别：

a) 实体类信息：指客观存在的、包含新闻实体公共特征属性的信息。如新闻事件、事发时间、主要人物等。

b) 业务类信息：指依据新闻领域内业务需求为导向的信息。如新闻场景、新闻背景、新闻情感等。

c) 多媒体类信息：指在新闻图片、音频、视频中存在的信息。如：图片场景、图片人物、图片属性、视频人物、视频场景、视频属性、地标建筑、音频人物、音频内容、其他。

5 实体类信息

5.1 实体类信息类型

实体类信息类型包括：

——新闻事件

——事发时间

——相关时间

——事发地点

——相关地点

——新闻主体
——主要人物
——相关人物
——主要机构
——相关机构

5.2 实体类信息详情

5.2.1 新闻事件

新闻报道中对新闻事实的概要性描述。

标注要求：新闻事件分为命名新闻事件和一般新闻事件。命名新闻事件从配备固定的受限词表中选取，一般新闻事件根据新闻报道内容开放式填写。可通过事件抽取算法实现自动标注。

示例： “庆祝改革开放 40 周年大会”“达沃斯论坛”

5.2.2 事发时间

新闻事件实际发生的时间。

标注要求：事件发生的当地时间。

示例：《习近平抵达印度金奈 出席中印领导人第二次非正式会晤》一文中提到“当地时间下午 2 时 10 分许，习近平乘坐的专机抵达金奈国际机场。”则事发时间为“当地时间 2019 年 10 月 11 日下午 2 时 10 分”。

5.2.3 相关时间

新闻事件事发时间之外，新闻报道中提及的时间。

标注要求：除新闻事件发生的时间以外提及的时间，要求不与发稿时间、事发时间重复。可通过自然语言处理算法实现时间要素的分析和自动标注。

5.2.4 事发地点

新闻事件实际发生的地点。

标注要求：从新闻报道中分析发生地的行政区划、地理位置、地标设施、相关数据。

a) 行政区划。指国家为进行分级管理而实行的区域划分。
b) 地理位置。指对新闻发生地点的地理信息的定量刻画。
c) 设施。指人为建造的并在此中进行相关活动的地点，包括建筑和交通设施及地标等。
d) 相关数据。指新闻事件发生时所处环境的实时基础数据和历史基础数据，主要由各类传感器采集，包括定位坐标数据、时间数据、设施数据、历史影像数据等，如设施扫描数据、卫星数据、历史图片与视频等。该类数据主要通过传感器技术采集，具有实时、连续、可比较、可解释与不可更改的特性。可用于新闻报道中实时记录、事后溯源、复现、分析、深度调查等的原始、基础依据，适用区块链、机器学习、无监督学习等技术对媒体挖掘与调查分析的辅助应用

示例1：地点为行政区划的，比如北京市海淀区、纽约市等
示例2：地点为地理位置的，比如北纬 35° 等
示例3：地点为地标设施的，比如欢乐谷、埃菲尔铁塔等
示例4：地点为相关数据的，比如由传感器采集、设施扫描的数据等

5.2.5 相关地点

非新闻事件实际发生的地点，新闻报道中提到的行政区划、地理位置，地标设施，相关数据。

标注要求：从新闻报道中分析提及的行政区划、地理位置、地标设施、相关数据，要求不与事发地点重复。可通过自然语言处理算法实现相关地点要素的分析和自动标注。

示例见5.2.4

5.2.6 主要人物

新闻发生涉及的主体人物。

标注要求：在新闻报道中占主要成分，常见于标题或导语中。需将主要人物的职务和姓名一同标注。可通过机器学习、自动标注等进行人物要素的分析。

示例： “美国总统特朗普表示我亲眼看到了非常了不起的中华文明和中国取得的非常伟大的成就。美方愿同中方达成彼此都可接受的贸易协议，这将具有历史意义。”主要人物为：美国总统特朗普。

5.2.7 相关人物

新闻报道中除主要人物以外的，作为出席、陪同形式出现的人物。

标注要求：新闻报道提及的人物，常见于新闻报道结尾部分。记者、编辑不列为相关人物。可通过机器学习、自动标注等领域人物要素的分析。

示例：《习近平在陕西榆林考察时强调 解放思想改革创新再接再厉 谱写陕西高质量发展新篇章》的结尾处有“丁薛祥、刘鹤、陈希、何立峰和中央有关部门负责同志陪同考察。”相关人物标注为“丁薛祥”、“刘鹤”、“陈希”、“何立峰”。

5.2.8 主要机构

新闻事件报道中起主要作用的组织机构。

标注要求：新闻事件的主要机构包括政府组织、军事组织、商业组织、非盈利行组织、医疗机构、教育机构等。主要机构描述常见于标题或导语。

示例：《公安部部署全国公安机关开展 2021 年烈士纪念日活动》中主要机构标注为“公安部”。

5.2.9 相关机构

新闻报道中除主要机构以外提及的其他组织机构。

标注要求：相关机构包括政府组织、军事组织、商业组织、非盈利行组织、医疗机构、教育机构等。不可与主要机构重复，可通过自然语言处理算法实现组织机构要素的分析和自动标注。

6 业务类信息

6.1 业务类信息类型

业务类信息类型包括：

—— 体裁
—— 国内/国际分类
—— 新闻分类
—— 摘要
—— 关键词
—— 新闻场景
—— 新闻背景
—— 原文标识
—— 策划类型
—— 新闻情感倾向
—— 政治术语
—— 引用（典）

6.2 业务类信息详情

6.2.1 体裁

新闻报道的表现形式。

标注要求：根据表达新闻的手法、口吻和组织材料结构的不同进行区分，包括：消息、通讯、评论、公文公报等。

a)“消息”报道事情的概貌，较为简短，内部无二级标题，宜500～800字以内。通过标题、导语、主体三层推进。

b)“通讯”运用叙述、描写、抒情、议论等多种手法，形象地反映新闻事件或新闻人物。通讯相比消息内容更长，且内部可存在多级标题。

c)“评论”是新闻传播机构发表的各种评论形式的报道。包括：述评、社论、评论员文章等文章。

d)“公文公报”指政策文章、领导人讲话稿、《求是》杂志发表的文章、公报、受权发布、党政机关和人民团体等授权媒体公开发布重大事件或重要决定事项的报道性公文公报。

e)其他体裁，指不属于上述类别中的体裁，如综述类新闻、回忆录、杂文等。

6.2.2 国内/国际分类

新闻报道所属的地域分类。

标注要求：根据新闻事件发生的地点进行分类，港澳台属于国内新闻。在中国国内发生的新闻事件，或者在公共海域完全由中国主导的新闻事件属于国内新闻。发生在中国以外的国家或地区的新闻属于国际新闻。

示例1：国内新闻如《上海市人民政府关于印发<上海市公有住房差价交换办法>的通知》

示例2：国际新闻如《一图读懂英国"脱欧"为何一脱再"拖"》

6.2.3 新闻分类

新闻分类代表新闻描述的主题。

标注要求：参考中文新闻信息分类与代码GB/T 20093或其他分类标准。可按照人物、组织等以及事件本身所属领域的相关度选择 1～3 项。

示例：《习近平出席亚运会开幕式》，标注为政治类和体育类新闻。

6.2.4 摘要

新闻报道内容的要点摘录。

标注要求：从标注新闻中摘取最主要的新闻元素，体现时间、地点、主要人物/组织、发生的事件，字数在100-150字左右。

6.2.5 关键词

新闻报道中的关键性内容，包括实体词、谓词、具有关键信息的词语。

标注要求：要求选择言简意赅，具有检索意义的词汇，有较特殊意义的词、词组、缩略语，不宜拆开。若新闻中有其他类别的词、短语甚至熟语也能够提示文章的关键内容，也应作为关键词处理。

6.2.6 新闻场景

新闻专题报道所属的场合、情景类型。

标注要求：涉及领导人的新闻报道标注相关场景，从制定的新闻场景类型表中选择填写，如国内视察、出国访问、会见、参会、出席重要场合等。

示例见表1：

表1 新闻场景类型

名称	举例
重要活动	会见、会晤、参观、视察
出国访问	会见、会晤，出发、到达、讲话
重要讲话	讲话、联合声明、其他
重要会议	党代会、全国人民代表大会、政治协商会议、研讨会、论坛、对话会、座谈会、专题讨论会、表彰会、全体会议
决定、命令、计划	主席令、嘉奖令、通令
重要文章	署名文章、讲话原文
函电贺词	致电、贺信
指示批示	

6.2.7 新闻背景

新闻消息稿中出现的一段对新闻中的人物、地点或者事件的扩展背景进行描述的内容。

标注要求：标记出新闻报道中与主体事件有解释性的、描述历史背景的文字内容。

6.2.8 原文标识

针对政策性的新闻报道，将包含有非转述的、原始的内容标记为原文，否则标记为非原文。

标注要求：通过原文标识对新闻文本材料进行区分，通过布尔值标记是否为原文。原文包括领导人讲话、工作报告、条例章程、谈话、白皮书、演讲、答问、批示、贺信、题词、署名文章、主旨讲话等。当新闻报道内容为第三人转述内容比如 XX 说，XX 指出，则不属于原文。

6.2.9 策划类型

描述新闻内容与当前新闻专题的关系。

标注要求：与新闻专题报道直接相关的为程序稿，与新闻专题报道不直接相关的，如周边报道、反馈稿件等为配合稿。

示例1：对事件内容进行烘托、背景资料等信息进行阐述的为配合稿。如《(习近平出访配合稿)背景资料：伊朗伊斯兰共和国》。

示例2：《年终特稿丨不忘初心 阔步前行》为年终稿。

6.2.10 新闻情感倾向

新闻报道中新闻主要人物或主要机构对新闻事件的感情、态度、意向或立场。

标注要求：

a)判断文中新闻主要人物、组织或机构对某人或事件表达的态度，感情倾向分为正面、负面、中性。

b)文中主要人物、组织或机构对多个事件对某人或事件表达的态度不同时，也需分事件将其标注出。

示例：新闻情感倾向类型的举例见表 2。

表2 新闻情感倾向类型与举例

类型	举例
正面	祝贺、庆祝、赞扬、表扬、感到开心、勉励、鼓励、感谢、积极评价、高度评价、热烈欢迎、祝福等
中性	正常陈述，无情感流露的
负面	愤怒、批评、指责、反对等

6.2.11 政治术语

新闻报道中文中出现的政治政策、政治口号、政治精神的表述或者缩写。

标注要求：

a)政治术语由连续或不连续的词语和短语整合而成;常在含义上表现出高度的凝固性、高度概括性;形式上较为简洁、凝练。

b)有三、四、五字等类似惯用语或成语形式的。

c)有呈对偶形式或成对、呈排比形式出现的。

示例：“不忘初心、牢记使命”、“两个维护”、“两个一百年”。

6.2.12 引用（典）

新闻报道中，被领导人引用的典故、熟语（成语、惯用语、歇后语、谚语）、古文诗词、格言警句。

标注要求：

a)领导人引用的典故，在形式上使用双引号“”标记出的。若未用引号标记，则不算做是引用。

b)熟语（成语、惯用语、歇后语、谚语)、古文诗词、格言警句，出现在双引号内部或“俗话说”、“古语言”等表述类动词之后的，均为用典。

示例：“人心所归，惟道与义”

7 多媒体元素类信息

7.1 多媒体元素类信息类型

多媒体元素类信息类型包括：

——人物元素
——场景元素
——地标建筑
——语音内容
——图片属性
——音频属性
——视频属性
——其他

7.2 多媒体元素类信息详情

7.2.1 人物元素

新闻图片、音频、视频等多媒体稿件中出现的人名、人脸，进行身份判断和标记。

标注要求：

a)对图片、视频、音频等多媒体稿件中出现的新闻人物身份进行判断、标记。

b)可使用人脸识别、语音识别等算法预处理人物信息，标注过程中需要进行人工确认。

7.2.2 场景元素

新闻图片、音频、视频等多媒体稿件中出现的新闻场景、情景分类。

标注要求：根据图片、音频、视频中所展现的内容、画面判断其所处的场景，进行标记。可使用机器视觉算法预处理的图片场景的信息，标注过程中需要进行人工确认。

7.2.3 地标建筑

新闻图片、音频、视频中出现的标志性的地理建筑。

标注要求：

a)根据新闻图片、音频、视频中出现的标志性的地理建筑进行标记

b)可通过机器视觉、语音识别、自然语言处理等技术，识别多媒体元素中所出现的地标、建筑物等地点。如故宫、自由女神像、泰姬陵等。标注中需要进行人工确认。

c)可通过激光、无人机扫描、多模传感器、卫星多时相、多通道、多谱段卫星遥感影像动态数据等实现对新闻图片、视频地标建筑的标注。（传感器新闻信息介绍参见附录A）

7.2.4 语音内容

新闻音频中的语言内容。

标注要求：可通过语音识别技术手段，将音频中的语音内容转化成文字记录。识别的语音结果需人工辅助核对。

7.2.5 图片属性

新闻图片的基本属性。

标注要求：通过读取图片的基本参数，获得基本属性，如横屏/竖屏，时间、位置、环境、像素、分辨率、大小、颜色、色调等。

7.2.6 音频属性

新闻音频的基本属性。

标注要求：通过读取新闻音频中的基本参数，获得基本属性，如音频时长、比特率、采样大小。

7.2.7 视频属性

新闻视频的基本属性。

标注要求：通过读取视频文件的基本参数，获得视频的基本属性，如横屏/竖屏、时间、位置、环境、时长、分辨率、码率、长度、宽度等。

7.2.8 其他

其他的多媒体元素标注标签。

标注要求：通过光学字符识别（OCR）、语音识别、机器视觉、自然语言处理、情绪识别等技术手段，识别图片、音频、视频中的文字、物品、受众情绪等内容，根据新闻检索需要增加的标签。

附 录 A
（资料性）
传感器新闻信息

在物联网时代，基于各类传感器采集的信息已成为新闻生产的重要来源。该类信息具备实时感知、多源多类、自动捕捉记录、可溯源、可解释、可印证和不可更改等重要特征。传感器采集信息源的技术是新闻报道中重要的数据挖掘辅助技术。

随着传感器深入到城市管理、智能交通、智慧生活和个人的健康管理领域，新闻报道采取传感器技术辅助报道已逐渐在业界推广使用。

传感器新闻在技术应用上主要分为物理传感和生物传感两大类。物理传感主要包括环境检测（水、气、土、农等）、遥感卫星（图像、视频、数据等，在此基础上可广泛实现对农业、工业、城市绿化、违建、重点标注进行分析）、无人机（图片、视频、网络中继等）、物理建模（激光等），一系列的物联传感已经广泛应用在各个行业，并建立了相关的行业标准，这些行业标准都可以作为标准引入到新闻信息的标注中。

生物传感器及感知智能技术涉及脑科学、神经科学、心理学、无线通信学、计算机科学、区块链等多学科和技术知识，可广泛应用于各类人机交互及多物种多样性场景，智能交通、教育、健康、文化等多领域。通过对受众的体验进行量化，测试受众在收看、收听内容时的情绪变化，对图片、音频、视频内容的分类、分级和改善用户体验均有较好的研究与使用价值，对用户的交互性和内容的质量评价具有科学性指导和深入研究的价值，丰富新闻报道的内涵、边界、深度和角度。