

《中文新闻信息结构化标注规范》编制说明

一、 标准的背景和依据

媒体融合是传媒业发展的顺势潮流；习近平总书记要求，要探索将人工智能运用在新闻采集、生产、分发、接收、反馈中，全面提高舆论引导能力。人工智能作为信息化时代背景下的热点技术，是促进媒体融合的重要推动力。美联社在 2005 年就开启了新闻信息结构化加工的进程，从最初开始研发新闻信息的分类法和规则，到逐步在内部实现基于规则的机器标注+人工辅助的自动化标注体系，再到 2012 年自动化标注系统正式对外服务，美联社利用自身的稿件资源进行深度挖掘，将分类应用于稿件生产的整个生命周期中，从而提高内部搜索的精准度，开发新产品，不断将稿件价值最大化。

目前，我国新闻领域在稿件数据深度挖掘、自动标注方面，仍处于起步阶段。建立中文新闻信息结构化标注规范，对推进媒体融合发展有着重要作用。如何在结合前沿技术的基础上，灵活运用这些数据创造新价值，通过规范化的信息结构化方法提供智能化服务是急需解决的问题。为此，新华通讯社通信技术局、北京语言大学、中国人民大学新闻学院等四家单位联合提出《中文新闻信息结构化标准规范》标准，为新闻信息资产的结构化处理提供指导性建议。

二、 标准的目的和意义

媒体融合需要内容和技术的双驱动，需要以先进技术为支撑、以优质内容为根本，充分发挥互联网思维优势，推动传统媒体和新媒体从简单的形式相加到深度的内容相融。新闻稿件数据是一笔宝贵的财富，在不断优化、升级新闻生产的过程中，离不开先进技术与稿件内

容的深度融合。随着新闻稿件数据量每年的增长，海量的非结构化稿件在一定程度上增加了技术处理的难度。对于人工智能技术来说，准确的学习样本能够带来精确的判断结果，规范新闻信息标注规则与方法，保障训练数据集标签的精准性和规范性，是实现技术突破和应用创新的重中之重。无论是通过人力手动标注，还是机器自动化标注，为保证标注的准确性，都需要一个具有科学性、实用性的规则规范来指导稿件数据标注，以减少缺乏行业专业知识等情况下造成的误差。因此，明确数据标注的规则和方法并将其标准化，对于保障新闻稿件数据的精准度来说具有重大意义。长期有助于产生高质量数据，优化数据内容结构，发挥海量数据潜在价值。在结构化数据的助力下，利用人工智能、移动网络及新兴媒体的技术和思维，加速新闻生产加工过程中的技术创新，更好地为记者、编辑们提供智能化服务，提升新闻采写效率。

三、 标准的编制原则

本标准参考中文新闻信息置标语言和中文新闻信息分类与代码国家标准，考虑到实际应用的科学性、实用性、统一性和规范性，对新闻信息的标注规则和方法给予指导和建议。本标准适用于新闻信息内容的标注环节，并可根据适用领域进行必要的扩充。

四、 标准编制过程

- 1) 标准主体部分由 4 家单位共同起草，各负责一个具体的章节。这样安排是为了行文的连贯性，避免起草方过多带来的沟通效率和内容碎片等问题；

2) 标准主体部分起草完成并合并成讨论稿后，交由所有参与方共同审核，讨论并改进；

3) 每一个章节均按照共识进行了简要描述并举例说明，根据现有内容对缺省内容进行补充使其完备；

按照上述思路，对编制的分工如下。“协助”的工作根据需要进行分配：

- 1) 前言：新华通讯社通信技术局起草，北京语言大学、中国人民大学新闻学院协助；
- 2) 编制说明：新华通讯社通信技术局起草，北京语言大学、中国人民大学新闻学院协助；
- 3) 新闻标注规范：新华通讯社通信技术局起草，北京语言大学、中国人民大学新闻学院协助；
- 4) 标准范围：新华通讯社通信技术局；
- 5) 规范性引用文件：根据实际引用由各起草单位提供，由星尘纪元智能科技有限公司协助汇总；
- 6) 术语和定义：根据实际引用由各起草单位提供，新华通讯社通信技术局、北京语言大学统一整理；
- 7) 层次结构目录：新华通讯社通信技术局起草，北京语言大学、中国人民大学新闻学院协助；
- 8) 新闻信息信息：新华通讯社通信技术局起草，北京语言大学、中国人民大学新闻学院协助；
- 9) 篇章整体属性层数据：新华通讯社通信技术局起草，北京

语言大学、中国人民大学新闻学院协助；

10) 新闻要素属性：新华通讯社通信技术局起草，北京语言大学、中国人民大学新闻学院协助；

11) 技术参数：星尘纪元智能科技有限公司起草；

标准编制经历意见征集、确立范围、编制分工、草案编写、合订意见、标准终稿等步骤，历时约1年，主要时间节点如下：

- 1) 2019年6月10日：标准编制筹备会议，成立标准编制工作组；
- 2) 2019年7月：完成标准编制说明初稿；
- 3) 2019年8月~2020年10月：标准编制意见征集期；
- 4) 2019年11月6日在新华社发行楼6层会议室召开启动会；
- 5) 2019年11月19日：标准编制分工说明；
- 6) 2019年12月~2020年3月：标准草案起草期；
- 7) 2020年4月20日：标准合订意见稿；
- 8) 2020年6月4日~9月30日：合订意见修改期；
- 9) 2020年10月26日：标准合订稿；
- 10) 2020年11月19日：在中国新闻技联标准化分会年会上，通过团体标准立项表决，并确定名称为《中文新闻信息结构化标注规范》；
- 11) 2020年11月20~2021年5月27日：根据立项意见建议完善标准内容，形成征求意见稿；

五、 本标准与其他标准的关系及协调性

本标准在编制过程中对以下标准进行了参考：

GB/T 20092-2013 中文新闻信息置标语言

GB/T 20093-2013 中文新闻信息分类与代码

截至目前，尚未发现本标准与我国有关的现行法律、法规和相关强制性标准相冲突。

六、 标准主要内容

本标准包含有如下几个主要内容：

- 1) 描述了新闻数据标注体系结构，阐明了中文新闻信息结构化标注的术语、定义
- 2) 规范标注规则、方法和要求；

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。使用对象包括媒体采编人员、内容运营或技术人员，新闻内容提供商及媒体应用软件开发和服务商。

七、 专利和知识产权情况说明

本标准不涉及专利和知识产权问题。

八、 标准性质的建议

鉴于本标准的内容，建议主管部门将该标准作为行业标准发布。

九、 遇到的问题和分歧及解决方法

在编写标准的过程中遇到的问题及解决办法主要体现在如下方面：

- 1) 新闻类别的划分：分类初期对于新闻体裁的区分需要进一步明确，包括分类名称、分类依据、对应结构等内容，确定了分类的树形结构之后可实现精准的打标。事件主体的表述：制定标准过程中，对于事件中的主体人物和相关人物的分类判别产生分歧，针对不同新闻场景，主体人物也不尽相同，特别是涉及国家领导人和大型会议事件的情况，经过协商和讨论，对于具体新闻情景的进行更加细致的判别，将人物参与事件的相关行为作为判别的重要依据，明确主体人物与相关人物的区别，实现准确的人物标注。
- 2) 术语和定义：初版草案关于新闻专业术语定义的学术性、专业性仍需进一步提升。由于早期标准的定义基本为工作组自行定义，部分虽参考了现行标准，但仍有部分术语不够精炼和准确。因此，工作组经协商参考 GB/T 20092-2013、GB/T 20093-2013，并针对本规范使用的范围进行一定的补充，保证了术语和定义的专业性。
- 3) 规范通用性：在早期的标准草案中，详细规定了新闻信息不同属性的数据类型。经工作组讨论认为，为满足不同场景下的数据标注需求，新闻时间、新闻人物、引用引典等不同属性的数据类型无需在规范中进行详细阐述和规定，以确保本标准的通用性和实用性；

- 4) 篇章结构：标准起草时，数据信息单独汇总在一章中。由于新闻信息中包含篇章整体属性、新闻要素属性、附加属性，与下面内容存在部分重复描述的情况，经讨论，将新闻信息根据不同属性类别进行拆分，作为二级标题添加至每个属性中，保证结构的清晰和完整。

十、 标准推广的措施建议

为了更好地发挥本标准的作用，建议采取以下措施推动贯彻实施：

- 1) 归口单位积极推动：标准发布后，由中国新闻技联牵头，组织召开标准宣贯会，通过媒体在相关产业链应用领域加强宣传，扩大标准的影响力。
- 2) 行业团体优先推广：考虑优先在相关产业链联盟组织等相关行业团体中推广实施，发挥行业团体的市场主体作用。
- 3) 在相关平台上提供免费下载标准文本服务。