ICS 01.020
CCS A 22

中华人民共和国国家标准

GB/T 44217.6—2024

语言资源管理　语义标注框架
第6部分：语义标注原则

Language resource management—Semantic annotation framework—Part 6：Principles of semantic annotation

［ISO 24617-6：2016，Language resource management—Semantic annotation framework—Part 6：Principles of semantic annotation（SemAF Principles），MOD］

2024-07-24 发布　　2025-02-01 实施

国家市场监督管理总局
国家标准化管理委员会　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。

本文件是GB/T 44217《语言资源管理　语义标注框架》的第6部分。GB/T 44217已经发布了以下部分：

——第6部分：语义标注原则；

——第11部分：可度量数量信息(MQI)。

本文件修改采用ISO 24617-6：2016《语言资源管理　语义标注框架　第6部分：语义标注原则(SemAF原则)》。

本文件增加了“规范性引用文件”一章。

本文件与ISO 24617-6：2016的技术差异及原因如下：

——更改了对数字语料库中的语义标注通常支持解释和推理的原因的阐述，将其改为陈述性描述(见6.3)，标准中无需阐述原因；

——删除了反馈关系和依赖关系未出现在话语关系标注方案的原因的阐述，删除了话语关系的ISO标注方案应从ISO 24617-2继承这些关系的原因的阐述(见ISO 24617-6：2016的7.4)，标准中无需阐述原因；

——更改了关于数值和量的标注方式(见9.3)，由于技术发展，对于数值和量的标注方式采用ISO 24617-11：2022中对数值和量的标注方式。

本文件做了下列编辑性改动：

——更改了标准名称，将语义标注框架的简称删去；

——更改了实现本文件目的三种方式的表述形式(见4.1)，根据GB/T 1.1—2020要求，改为列项表示；

——更改了示例的表述形式(见4.2)，改为符合汉语表述方式的示例；

——更改了示例1、示例2、示例3的内容(见4.2)，改为符合汉语表述方式的示例；

——更改了标注方案间一致性和标注方案集合完整性的表述形式(见第5章)，根据GB/T 1.1—2020要求，改为列项表示；

——更改了语义标注两个功能的表述形式(见6.3)，根据GB/T 1.1—2020要求，改为列项表示；

——更改了示例的表述形式(见6.3)，改为符合汉语表述方式的示例；

——更改了对有意义的标注的解释方式(见6.3)，标准中无需阐述原因，改为陈述性描述；

——更改了示例2的内容(见6.3)，改为符合汉语表述方式的示例；

——更改了示例4、示例5、示例6的内容(见7.2)，改为符合汉语表述方式的示例；

——更改了示例的表述形式(见7.2)，改为符合汉语表述方式的示例；

——更改正文中对于参考文献的表述方式(见7.2)，以符合GB/T 1.1—2020的要求；

——更改了给定表示转换为另一个语义等同表示的步骤的表述形式(见7.3)，改为列项表示；

——更改正文中对于参考文献的表述方式(见7.4)，以符合GB/T 1.1—2020的要求；

——更改了示例4中对距离的描述(见8.2)，改为符合汉语习惯的衡量距离的单位；

——更改了ISO对话行为标注方案两种类型的表述形式(见8.4)，改为列项表示；

——更改了示例1及相关内容(见8.8)；

——更改了示例的表述形式(见9.2)，改为符合汉语表述方式的示例；

——更改了示例内容及表述形式(见 9.4),改为符合汉语表述方式的示例;

——删除了原示例(24)及正文中相应的说明(见 9.5.1,ISO 24617-6:2016 的 8.5.1),其表述方式不符合中文习惯;

——更改了示例内容及表述形式(见 9.5.2),改为符合汉语表述方式的示例。

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。

本文件由全国语言与术语标准化技术委员会(SAC/TC 62)提出并归口。

本文件起草单位:中国标准化研究院、华南师范大学、湖北省标准化与质量研究院、北京信息科技大学、厦门坤锦电子科技有限公司、中国科学技术信息研究所、中国质量标准出版传媒有限公司、北京集贤弘文文化传媒有限公司、聊城大学、北京工业大学、上海对外经贸大学、中国中医科学院中医药信息研究所、江苏科技大学、四川语言桥信息技术有限公司、广州智语信息科技有限公司。

本文件主要起草人:曹馨宇、王海涛、郝天永、陈炎明、吕学强、黄树福、刘耀、刘晓东、严可、贺莉丽、魏洁、鲁曦、贾仰理、徐术坤、刘磊、刘亮亮、周洪伟、刘嘎琼、朱宪超、瞿瑛瑛。

引　言

语义标注是计算机对自然语言深层次处理的重要技术之一，是对文本中的词语或句子添加可供理解的语义标签的过程。依据标注的一般原则和具体标注对象的不同，标准被划分为不同的部分，GB/T 44217《语言资源管理　语义标注框架》拟由12个部分构成。

——第1部分：时间和事件。目的在于提供一种通用的方法来描述文本中的时间和事件。

——第2部分：对话行为。目的在于提供一种表示对话行为的标注语言以及一种将对话分割为语义单元的方法。

——第4部分：语义角色。目的在于为语义角色提供一个协商一致的标注方案。

——第5部分：篇章结构。目的在于为话语实现和话语内容提供一种表示方式。

——第6部分：语义标注原则。目的在于确定以语义标注框架为特征的语义标注方法。

——第7部分：空间信息。目的在于提供一种通用的方法来描述自然语言文本中表达运动相关的空间信息和时空信息。

——第8部分：篇章中的语义关系，核心标注框架。目的在于为话语关系的表示和标注提供一个方案。

——第9部分：引用标注框架。目的在于为自然语言文本和多模态交互中所指现象的标注和表示提供一个综合模型。

——第11部分：可度量数量信息(MQI)。目的在于为可度量数量信息提供一种标注方案。

——第12部分：数量。目的在于为数量信息语义表示提出一般形式化定义。

——第14部分：空间语义。目的在于通过为抽象语法建立形式语义提供标注空间信息的方法。

——第15部分：可度量数量信息抽取。目的在于提供一种从自然语言文本中抽取可度量数量信息的一般方法。

语言资源管理　语义标注框架
第6部分:语义标注原则

1　范围

本文件描述了以ISO语义标注框架(SemAF)为特征的语义标注方法。SemAF策略可为特定类别的语义现象开发独立的语义标注方案,并最终合成一个单一、连贯且覆盖广泛的方案,本文件对这一策略进行了简要叙述,并针对ISO语言标注框架中的"标注"与"表示",分别给出了用于语义标注的抽象句法概念与具体句法概念。本文件还描述了上述概念在元模型规范和标注语义解释方面的作用,以便确定一个理据充分的标注方案。

本文件还为SemAF各个部分的标注方案提供了指南,用于处理以下两个问题:一是因标注方案重叠而引起的概念与术语上的不一致;二是涉及多个SemAF部分的语义现象(如否定、情态和计量)的处理方式。本文件对以上问题均给出了确切实例,并视情况给出了部分的解决方案。

本文件适用于为不同语义现象设计标注方案。

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。

ISO 24617-2　Language resource management—Semantic annotation framework (SemAF)—Part 2:Dialogue acts

3　术语和定义

下列术语和定义适用于本文件。

3.1

原数据　primary data

文本或交流行为的电子化表示。

示例:文本的数字表示、语音转录、手势或多模式对话。

注1:ISO 24612将原数据定义为"语言数据的电子表示"。对于本文件,这个定义并不太合适,因为语义标注也可以与非语言或多模态数据有关,例如带有伴随手势和面部表情的口语对话,甚至是没有任何伴随语言的手势和/或面部表情。

注2:原数据指未进行标注的原始数据。

3.2

标注　annotation

添加到原数据(3.1)的与其表述无关的语言信息。

[来源:ISO 24612:2012,2.3]

3.3

语义标注　semantic annotation

包含与原数据(3.1)片段或区域的含义有关的信息的标注(3.2)。