2000-2023 年《情报学报》全文内容细粒度研究方法实体标注规范

郝家亿[1]，王玉琢[2]，章成志[1]

1. 南京理工大学经济管理学院，南京 210094
2. 安徽大学管理学院，合肥 230601

本部分对研究方法实体的标注规范进行介绍，主要包括具体标注示例及其他注意事项。

1 研究方法实体标注具体说明

1.1 研究方法实体类型划分

依据正文对情报学领域的细粒度研究方法实体分类，在标注时笔者以句子为单位，若一个句子中包含研究方法实体（“理论实体”/“单一型方法实体”/“复合型方法实体”/“数据集实体”/“指标实体”/“工具实体”/“其他实体”），则对该句中出现的所有研究方法实体进行标注。为了标注方便，笔者将上述类别分别用以下形式进行标注：“theory”、“single method”、“compound method”、“dataset”、“metrics”、“tool”、“others”。本文借助在线标注平台 INCEpTION 进行人工标注，标注界面如图 1 所示

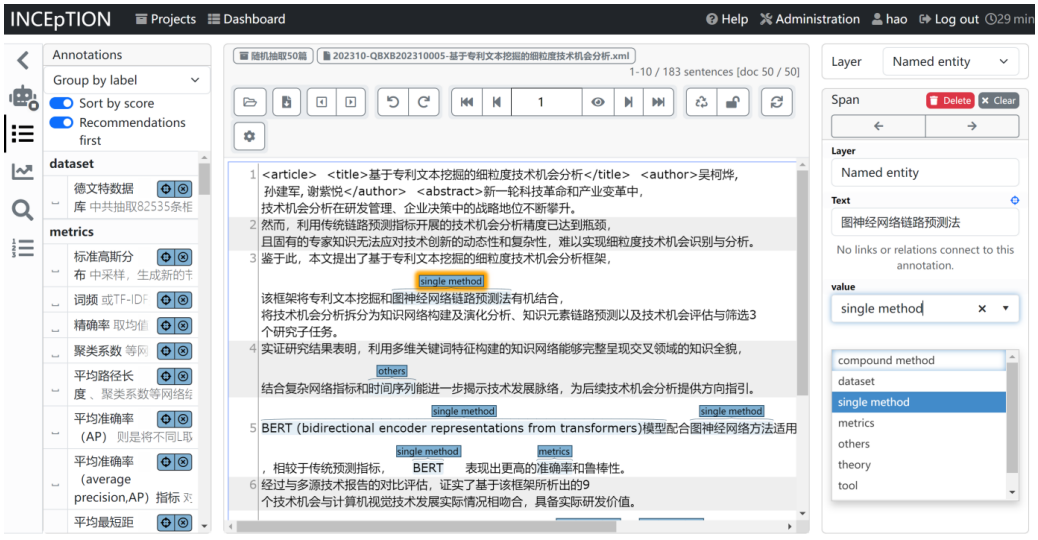

图 1： INCEpTION 工具标注界面示例

1.2 研究方法实体标注注意事项

- **限定词**（置于名词前起限定作用，如其他的、本文的等）不应成为实体范围的一部分。如“本文提出的评论自动摘要方法”，仅标注“评论自动摘要方法”为单一型方法实体。

- **最小范围原则**：标注人员只需标注表示理论/单一型方法/复合型方法/数据集/指标/工具/其他实体的原始含义所需的最小范围，例如不是研究方法实体本身必要组成成分的定语修饰语（一般是形容词），不需要标注。

表2：最小范围原则示例1

例句	注释	实体类型
利用向量间的互补余弦值来表示论文之间研究主题的相似程度 。	互补余弦值	指标实体

如表 2 所示，针对短语“向量间的互补余弦值”，只标注“互补余弦值”为指标实体，前面的定语“向量间的”则不需要标注。

表3：最小范围原则示例2

例句	注释	实体类型
传统的 CA 模型在研究知识交流问题时，每个元胞表示一个知识主体，由于本文考虑不同地理空间知识主体的交流， 因此，将元胞 C 定义为聚集了大量知识主体的地理空间。	CA 模型	单一型方法实体

如表 3 所示，针对短语“传统的 CA 模型”，只标注“CA 模型”为单一型方法实体。

表4：最小范围原则示例3

例句	注释	实体类型
在数据获取方面，用户可以通过网页和 API（application programming interface）等方式访问并获取 Dimensions 数据库资源，以开展大规模数据分析 。	API; application programming interface; Dimensions 数据库	工具实体；工具实体；数据集实体

如表 4 所示，针对短语“Dimensions 数据库资源”，仅标注“Dimensions 数据库”为数据集实体。

- **最小范围原则（续）**：但如果是模型之间的对比，则需要根据上下文判断是否保留研究方法实体前后的修饰语。

表5：最小范围原则示例4

例句	注释	实体类型
深度迁移学习常用的方法包括基于实例的深度迁移学习	深度迁移学习；基于实例的深度迁移学习；	单一型方法实体；复合型方法实体；

（instance-based deep transfer learning）、基于映射的深度迁移学习（mapping-based deep transfer learning）、基于神经网络的深度迁移学习（networkbased deep transfer learning）以及基于对抗的深度迁移学习（adversarial-based deep transfer learning）。	instance-based deep transfer learning； 基于映射的深度迁移学习； mapping-based deep transfer learning； 基于神经网络的深度迁移学习； networkbased deep transfer learning； 基于对抗的深度迁移学习； adversarial-based deep transfer learning	复合型方法实体； 复合型方法实体； 复合型方法实体； 复合型方法实体； 复合型方法实体； 复合型方法实体； 复合型方法实体

如表 5 所示，针对短语“基于实例的深度迁移学习”，如果去掉其中的定语修饰语“基于实例的”，跟前面的实体“深度迁移学习”就没有区别了，因此这里的修饰词就需要保留。

表6：最小范围原则示例5

例句	注释	实体类型
同样地，为了充分学习图片中与文本实体相对应的语义特征，Yu 等提出基于 Transformer 架构的多模态命名实体模型，该模型主要由单模态特征表示、多模态 Transformer 及辅助实体边界检测组成，通过这些构件，模型能够较好地学习文本和图片上下文敏感特征，并能够关注到聚合多模态信息时未被充分关注的实体。	基于 Transformer 架构的多模态命名实体模型	复合型方法实体

如表 6 所示，针对短语“基于 Transformer 架构的多模态命名实体模型”，保留其中的“基于 Transformer 架构的”，因为该模型经语义介绍包括多模态 Transformer，故都需要标注进行区分。

表7：最小范围原则示例6

例句	注释	实体类型
本文通过模型输入拟人化、卷积神经网络（convolutional neural network）编码以及融合门机制并结合长短时记忆单元（long short-term memory, LSTM）优化了语言模型，提出了结合 LSTM 和 CNN 混合架构的深度神经网	卷积神经网络； convolutional neural network； 长短时记忆单元； long short-term memory； LSTM； 结合 LSTM 和 CNN 混合架构的深度神经网络语言模型 ；	单一型方法实体； 单一型方法实体； 单一型方法实体； 单一型方法实体； 单一型方法实体； 复合型方法实体； 复合型方法实体

络语言模型 (Gated CLSTM)。 Gated CLSTM

如表 7 所示，对于短语“结合 LSTM 和 CNN 混合架构的深度神经网络语言模型 ”“LSTM”和“CNN”都属于方法实体，“深度神经网络语言模型”是结合二者提出的一种复合型方法，故将整体标为一个复合型方法实体。

- **匿名实体。**不要标注包括照应语[①]的匿名实体。

【示例】“算法一”、“整个数据集”、“这个模型”等不需要标注。

- **具有特定名称的实体。**单一型方法/复合型方法/工具/数据集/指标/理论/其他实体通常具有特定的名称，并且其含义在不同的论文中是一致的。本文仅注释“事实，具有内容的”实体，特别要注意标注含义在不同论文中是一致的实体，这些实体一般是单个字符、词组或者短语，如果其是“事实、具有内容的”实体然而表达的含义仅局限在一篇论文的主题或研究问题中，也不需要标注。

【示例 1】

基于此，本研究同样通过构建学科向量并计算学科相似度，来衡量科学基金项目与产出论文在学科分类上的目标一致程度。其中“学科相似度”表达的含义局限于当前论文的研究主题，故不需要标注。

【示例 2】

在梳理范畴间逻辑关系的基础上，构建出多源数据驱动下产业竞争情报智慧服务机制作用模型。

其中“产业竞争情报智慧服务机制作用模型”虽然具有特定含义，但其仅限于当前论文，不需要标注。

【示例 3】

“本文具体使用的文档表示学习方法是Doc2Vec，该方法是Mikolov等基于 Word2Vec 模型提出的一种神经网络语言模型”，这里的“文档表示学习方法”和“Doc2Vec”都需要标注。

① 照应语，语言学术语。“先行词”的对称。指语言片段中照应先行词的语言成分。一般由人称代词和指示代词充当。The book which I borrowed is now overdue, I shall take it back to the library tomorrow. 其中 which 和 it 是 the book 的照应语。

- **缩写。**如果句子中同时存在全名和缩写，则将缩写及其相应的全名分别注释。

【示例 1】

“最后，使用层次狄利克雷过程（hierarchical Dirich‐let process,HDP）模型提取文本主题 。”

其中将“层次狄利克雷过程”、“hierarchical Dirich‐let process”和“HDP”标注为单一型方法实体。

【示例 2】

“其次，CPP 表示论文在 Web of Science (WoS)中的篇均被引量。”

对其中“Web of Science”和“ WoS”分别标注为数据集实体。

- **尾名词。**方法实体末尾如果出现 “算法”、“框架”、“模型”、“数据库”等词语，则一并标注。例如，方法实体“LDA”在一篇论文中可能有“LDA 算法”和“LDA”两种表示，这个时候如果实体名称后面跟有提示词“算法”则一并标注，如果没有提示词则只标注“LDA”。

【示例 1】

数据集“人民日报语料库”、“宾州树库”；算法或者模型“PLDA 模型”、“DTM 模型”、“LDA 主题聚类”等实体名称后面跟的提示词都需要标注。

表8：尾名词示例2

例句	注释	实体类型
人民日报语料库本文使用的是 1998 年 1 月的人民日报语料库	人民日报语料库； 1998 年 1 月的人民日报语料库	数据集实体； 数据集实体

如表 8 所示，这里的例子比较特殊，同一个数据集实体的两种表述都要标注，因为后者是前者的进一步阐述，因此后者需要将限定词“1998 年 1 月的”加入。

- **统称。**某类方法实体的统称不需要标注，只标注有具体名称和确定含义的实例。例如“分类算法”、“非监督学习”、“机器学习”、“深度学习”等不需要标注，只标注具体的方法实体如 CNN、RNN、Transformer、SVM 算法、决策树等。

【示例】

“把公式（1）抽象为语料层、文本层、词语层。利用图模型的方式把 LDA 模型表示出来，如图 2 所示。”然而，这里的“图模型”则需要标注为单一型方法实体，因为图模型是一种较为常见的表达。

- **连词。**如果一个连词（如“或”、“和”）连接的名词是一个整体，则标注整个短语；否则，分别单独标注每个实体名称。

表9：连词示例1

例句	注释	实体类型
石春丹等提出利用双向门控循环网络与 CRF 结合的模型对文本中人名、地名和机构名等实体进行识别，该模型能够有效学习序列的时序信息，并能捕捉长距离依赖 。	双向门控循环网络与 CRF 结合的模型	复合型方法实体

- **指标连词。**注意只限于本文的指标实体暂不标注。

表10：指标连词示例1

例句	注释	实体类型
Kruskal-Wallis 检验和逻辑回归的方法，研究了正面引用和中性引用论文在影响力上的差别引用原因对正面引用论文的影响力的作用，并进一步探讨了施引位置、引用长度参考文献数以及引用强度等因素对正面引用和中性引用的影响。	Kruskal-Wallis 检验； 逻辑回归	单一型方法实体； 单一型方法实体

在上述例句中，“施引位置、引用长度参考文献数以及引用强度”属于作者自行定义的指标，不用标注。

- 注：规则类词汇不属于研究方法实体，对此不进行标注，如“基于 OCC 模型从网民认知角度建立情感规则”中的“情感规则”不属于这六类中的任何一类。

2 其他注意事项

本研究使用的是 XML 格式的论文全文语料，由 html 格式的论文转录而成，但在数据转换的过程中会出现一些错误，例如可能会出现以下三种情况：

- html 格式转成 xml 时数字的下标和上标是没有的。如“从多项式分布 zi 中抽样 ωi，得到 P（ωi|zj, β），”，变为“从多项式分布 θ 中抽样 Z，得到主题概率 P(z | α) 。”

- 由于格式规范问题，正文中的“1）”及其以后的小标题并不会出现。如“1）技术主题提取 LDA 主题模型能够反复迭代自发学习进化，识别大规模文档集中潜藏的主题信息”中，“1）技术主题提取”该标题不会识别。

- 经转化后原文“D&M IS Success 模型”中的“&”会变为“&”。