

桥接指代标注中的主观性

Lauren Levine and Amir Zeldes

Georgetown University

Department of Linguistics

{ lel76, amir.zeldes } @georgetown.edu

Abstract

衔接指的是话语中可推理实体与其前件之间的关联关系，这种关系帮助我们理解这些实体，例如理解一个提到的“房子”时“门”是什么意思。由于识别实体之间的关联关系本质上是一个主观任务，因此在注释桥接回指及其前件时很难实现一致的协议。在本文中，我们探讨在三个层次上桥接实例注释中涉及的主观性：回指识别、前件解析和桥接子类型选择。为此，我们对现有的 GUM 语料库的测试集进行了一项注释试验，并提出了一种新开发的桥接子类型分类系统，并与以前提议的方案进行了比较。我们的结果表明，一些以前的资源可能严重注释不足。我们还发现，尽管在桥接子类型类别上的一致性中等，但在穷尽识别桥接实例方面，注释者之间的重叠较低，许多分歧源于对涉及实体的主观理解。

1 介绍

语篇连贯是一种指示现象，其中新引入的语篇实体在解释时依赖于相关但非相同的前文实体。术语“语篇连贯”指的是语篇参与者从他们当前正在处理的实体回溯到前文实体 (Clark, 1975) 的推测构建。这种关联关系可以由多种语言机制触发，包括词汇的部分-整体关系（如房子 - 门）和隐含的论元（如谋杀 - 受害者）。自从 Clark (1975) 首次评论这种现象以来，它收到了多种理论处理，包括 Prince (1981) 提出的与之密切相关的可推导概念，该概念将信息状态作为识别指示性语篇连贯关系的关键组成部分。这种理论上的分歧导致了一些不同的标注形式，其中在语篇连贯的定义以及语篇连贯的子类划分上均有所不同。尽管最近针对跨不同语料库的语篇连贯标注进行了一些协调努力 (Levine and Zeldes, 2024)，目前语篇连贯资源的总体状况仍然是异质的。语篇连贯资源的内部和之间缺乏一致性主要源于它们对语篇连贯的不同定义，以及在识别语篇连贯实例时主观注释者的判断。

在本文中，我们探索连接照应标注中的主观

性，以理解如何考虑这种主观性并在未来的工作中创造更一致的标注。我们考察了标注过程中三个阶段，标注者在这些阶段需要做出主观判断：（1）识别连接照应词，（2）解析回其相关的前项，以及（3）识别连接对的子类型类别。为此，我们在现有英语语料库 GUM (v10) (Zeldes, 2017) 的测试集上进行了标注试验。虽然 GUM 语料库包括连接标注，但标注指南不详且不包括连接子类型标注。此标注试验是开发新的连接资源 GUMBridge 的初步阶段。为了这一工作，我们开发了一个新的连接子类型分类系统，该系统以三种关系类型组织：COMPARISON 关系、ENTITY 关系和 SET 关系，以及一个额外的 OTHER 类别。我们还创建了标注指南，说明如何识别连接照应词-前项对的实例，以及如何将它们分类为子类型。

通过分析这次试点的结果，我们发现一方面我们能够识别出比以往几个资源所建议的更多和更密集的桥接佐证。从主观性角度看，我们发现对于桥接子类别的选择以及给定照应词的前项选择有中等程度的一致性。然而，尽管多数情况下准确性合理，桥接照应的识别中注释者的重叠度却显著较低。我们对试点中的注释进行定性评估，发现主观性在上述三个注释者判断阶段中都起着作用，尤其是在召回率上。我们探讨了主观性在每个阶段的作用，然后给出了关于如何构建桥接照应注释的建议，以考虑到注释者判断中的主观性。

2 背景

如上所述，用于桥接的注释形式主义有很多种，对桥接作为一种现象的定义也有所不同。在英语中，桥接解析系统（旨在自动识别桥接照应并解析回其关联先行词的系统）的评估通常使用以下三个语料库：ISNotes (Markert et al., 2012)、BASHI (Rösiger, 2018) 和 ARRAU RST (Poesio and Artstein, 2008; Uryupina et al., 2019)。ARRAU RST 通过标识建立文本连贯性的提及对来注释桥接实例，然后通过预定义的一组语义关系进行分类；而 ISNotes 和 BASHI 则基

于实体的信息状态来注释桥接照应，认为桥接是媒介信息的一种子类。

实体的信息状态 (IS) 指的是该实体对于话语的读者/听者的可接触程度 (Nissim et al., 2004)。一般来说，“新”信息是读者/听者无法识别的，而“已知”信息是可以识别的。“已知”实体可能由于多种原因被读者/听者识别：实体可能在先前的话语中已经被介绍（共指），实体可能通过通用知识/世界知识可被接触，或者在关联的情况下，实体的指称可以从先前的话语实体中推断。关联和通用知识/世界知识的实例都被视为“可接触的”，因为它们在首次被引入话语时就被读者/听者所识别，但只有关联的实例依赖于联想性前件来理解。

	Tokens	Bridging Instances	Bridging per 1k Tokens
ARRAU RST	229k	3.7k	16.5
ISNotes	40k	663	16.6
BASHI	58k	459	7.9
GUM (v10; full)	228k	1.9k	8.3
GUM (v10; test only)	26k	222	8.5
GUMBridge (v0.1)	26k	401	15.4

Table 1: 几个英语桥接资源的桥接实例的频率。

还有许多其他现有的桥接资源：在英语中，有 GUM、SciCorp (Roesiger, 2016)、corefpro (Grishina, 2016)、RED（更丰富的事件描述，O’Gorman et al. 2016）；以及其他语言中：德语中的 GRAIN (Schweitzer et al., 2018) 和 DIRNDL (Eckart et al., 2012)，捷克语中的 PDT (Nedoluzhko et al., 2009)，波兰语中的 PCC (Ogrodniczuk and Zawisławska, 2016)，仅举几例。此外，在与桥接密切相关的领域也有一些努力，例如 Recasens et al. (2010) 提出了一种用于分类共指的近同一性关系 (NIDENT) 的类型学，以及 Modjeska (2004) 关于其他照应的研究，我们现在将其视为桥接的一种副类型。我们提供了关于 ISNotes、BASHI 和 ARRAU RST 的背景信息，因为它们常用于桥接解决评估 (Yu et al., 2022; Kobayashi et al., 2023)，并且它们展示了识别桥接实例的不同视角。表 1 显示了这三种资源、原始 GUM 桥接注释和本文描述的 GUMBridge 注释试点中产生的桥接注释的比较统计数据。

ISNotes 是一个由 OntoNotes 语料库 (Weischedel et al., 2011) 中的 50 篇《华尔街日报》(WSJ) 文档组成的语料库，注释了细粒度的信息状态。ISNotes 区分了信息状态的三个主要类别：New、Old 和 Mediated。Old 信息是听者已知的信息和/或之前提到过的信息，而 New 信息是首次引入的信息。Mediated 信息之前没有引入，但无法独立理解，需要从先前提

及或通用/现实世界知识中推断。在 Mediated 类别中，有六个子类别，包括 bridging。语料库中在 mediated/bridging 类别下包含 663 个桥接实例，并且在 mediated/comparison 类别中还有另外 253 个比较指代实例，这被认为是桥接的一种（~ 每 1000 个词就有 16.6 个桥接实例）。Markert et al. (2012) 报告了标注员对的 Cohen’s κ ，mediated/bridging 的范围为 ~0.6-0.7，mediated/comparison 为 ~0.8。他们指出，相对于其他信息状态类别，mediated/bridging 的一致性更依赖于标注员。

在 BASHI 语料库中，50 个来自 OntoNotes 语料库的 WSJ 文档上增加了注释，其中包括了总计 459 对搭桥配对（~ 每千个词 6.9 个搭桥实例）。Rösiger (2018) 引入了指称搭桥与词汇搭桥之间的对比，其中指称搭桥是一种真正的照应关系（指代的解释需要前件），而词汇搭桥则是两个实体之间的非照应语义关系。该语料库具体只包含指称搭桥的注释，不包括词汇搭桥。BASHI 中的搭桥实例具有定式、不定式和对比照应的子类型。这些类别的注释者一致性分别和总体都有报告。识别搭桥配对的总一致性为 59.3%，其中对比照应的一致性较高，为 71.4%，而定式为 63.8%，不定式则较低，为 42.3%。

ARRAU 是一个多体裁的语料库，涵盖了各种指代现象，由四个子语料库组成，每个子语料库都有其自己的注释规范。ARRAU RST 是最大的子语料库，也是用于桥接解析评估最多的一个。它由 WSJ 新闻数据组成，包含 3,777 个桥接注释（每 1,000 个标记有 ~16.5 个桥接实例）。ARRAU 的桥接注释通过非同源性关系连接相关的提及，以建立“实体连贯性”，但由于范围非常广，注释仅限于一组固定的语义关系。该语料库使用 9 种桥接子类型的列表进行注释：possession、element-set、subset-set、标记为 ‘other’ 的回指，以及前述的伴随逆关系，还有一个额外的 under-specified 关系。ARRAU 的桥接注释模式和指南是从 GNOME 项目 (Poesio, 2004) 扩展而来。在 GNOME 项目中，编码者在使用其固定关系集合选择桥接子类型标签上表现出高度一致 (95.2%)，但在一致识别桥接实例上召回率较低 (22%)。

将标注限制在一组预定义的关系中会限制桥接作为一种现象的范围，但也旨在提高标注的一致性。然而，如 Rösiger (2018) 中所指出的，从预定义的关系中标注也可能引入误报，即在语义关系的一个实例实际上并不是构成指称桥接的关联指代的情况下。例如，欧洲-西班牙的情况显示出一种部分关系，但它不是指代的，因为不必参考欧洲就可以理解西班牙。基于信息状态标注的观点旨在避免这种误报，与“实

体连贯性”的概念相比，它提供了更具体的语言学标准来识别桥接实例，并消除了因范围原因而仅标注预定义关系集的需要。然而，这种基于信息状态的方法也极大地扩大了应被视为桥接的范畴，这又增加了标注者主观判断的影响。因此，为了推进基于信息状态的标注视角，我们必须开发处理其产生的额外主观性的方法。

如同在表 1 中所看到的，以前的资源中桥接标注的频率有相当大的差异，其中 ARRAU RST（统计词汇和指示性桥接）和 ISNotes 识别桥接实例的频率每千个词大约是 BASHI 和 GUM v10 的两倍。这表明像 BASHI 和 GUM 这样的一些以前的桥接资源可能在桥接实例的标注上不足，并促使需要重新审视桥接标注程序。

3 注释试点

本文在标注桥接实例时对主观性进行了分析，使用的是一个标注试验的结果，该试验旨在创建一个名为 GUMBridge 的新桥接资源。GUMBridge 建立在 GUM 的基础上，GUM 是一个现有的多文体英语语料库，GUMBridge 旨在结合当前现有形式的各个方面：使用信息状态知情的方法来识别桥接实例（如 ISNotes 和 BASHI），然后通过语义关系的分类体系进行子类型分类（如 ARRAU）。此外，GUMBridge 还旨在为核心英语桥接资源增加文体多样性，因为 ISNotes、BASHI 和 ARRAU RST 都由 30 多年前的 WSJ 新闻数据组成，在文体多样性方面提供的分析较少。虽然这一资源的开发仍在进行中，但本文发布了 GUMBridge 测试集（版本 0.1）的桥接标注的仲裁版本¹。该仲裁过程的详细信息在第 3.5 节中描述。识别桥接实例的指南（v0.1）在第 3.1 节中描述，桥接子类型的分类系统（v0.1）在第 3.2 节中描述。

3.1 识别桥接实例

在 GUMBridge 的注释工作中，我们采用依据信息状态的视角来识别桥接回指的实例。如 2 节所述，一个实体的信息状态是指在其引入时该实体对读者/听者的可及程度。我们说一个实体是“可及的”，如果它之前没有被提到过，但其指代对于读者/听者是可推断的。当一个实体的首次提及是通过来自话语中先前的、非相同实体的推理“可及”时，就发生了桥接。与那些由于是泛指、世界知识的一部分或话语情境而可及的实体不同，桥接回指本身不是可及的，而是依赖于先前的实体进行解释。注释人员会被提供关于这种桥接和可及性的定义概

述，并被指导在判断某一特定实体是否为桥接回指时考虑以下几点：

1. 您是否认为这个实体在话语中在某种程度上是可接触的？
2. 该可及性是否依赖于对话中先前实体的理解？如果是这样，请确定该先前实体的最新一次提及。

如果实体符合上述标准，则它是一个桥接指示词，上一个实体是其关联前指。一旦识别出桥接对，就可以根据下一节中描述的内容为其分配一个子类别。

3.2 桥接亚型的分类

为了对 GUMBridge 中存在的桥接种类进行分类，我们创建了一个新的桥接子类型分类系统。该分类系统由 11 个类别组成，其中 10 个类别被组织在 3 种关系类型下：COMPARISON 关系、ENTITY 关系和 SET 关系，另外还有一个 OTHER 类别。为 GUMBridge (v0.1) 开发的桥接子类型分类系统如图 1 所示。下面是每种桥接子类型的简要说明。对 ARRAU 的桥接子类型的简要比较列在附录 C 中。

COMPARISON-RELATIVE 回指词之前常有比较标记（其他、另一个、相同、更多等）、序数（第二、第三等），或比较级形容词（更大、更小等），这暗示着与前件进行比较（或反之亦然）。

Several women 走进了房间。其他妇女随后跟进。

COMPARISON-TIME 指代词指的是一个特定的时间/时间框架，该时间/时间框架可以通过与前件所表达的时间/时间框架的参照来理解（或反之亦然）。

我去购物了 Wednesday, March 3rd。我将在接下来的星期三再次去。

指代词的类型被省略，但可以通过与先行词进行比较推断出来（或反之亦然）。

我去过 Chinese restaurant。我想去意大利的那个。

ENTITY-ASSOCIATIVE 回指项是先行词的一个属性或紧密关联的实体（或反之亦然）。这通常表现为例子 3.2 中谓词的隐含论元，例子 3.2 中的关系名词，以及例子 3.2 中的典型关联：

昨晚发生了 a murder。受害者尚未被确认。公园里有 a child。它的父母一定在附近。我上周去了 a wedding。接待会非常有趣。

¹<https://github.com/lauren-lizzy-levine/gumbridge>

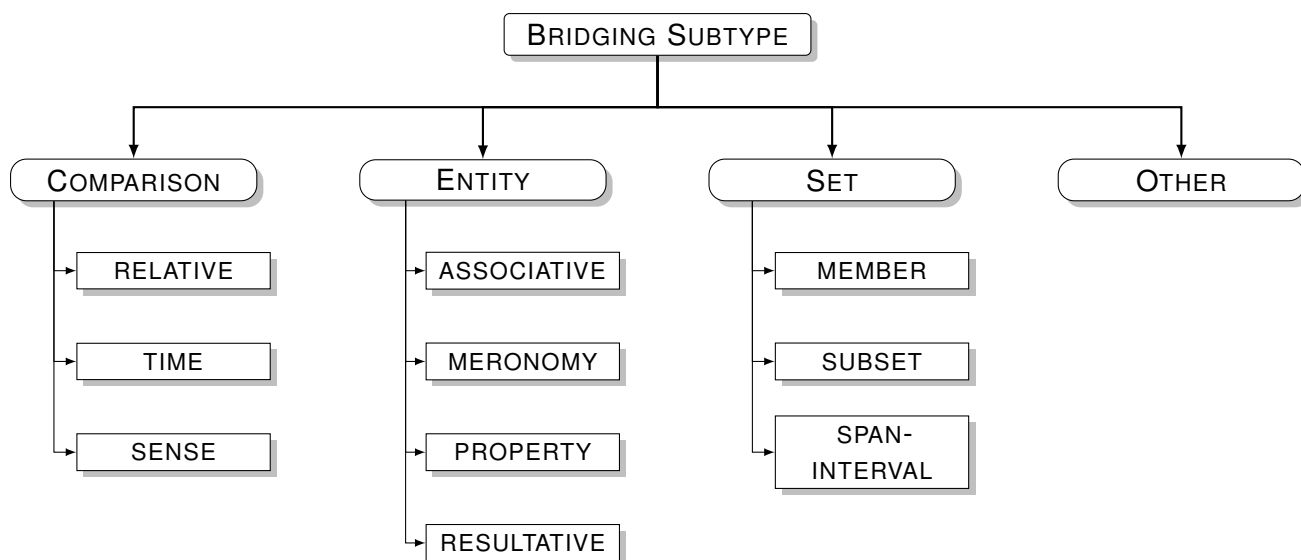


Figure 1: GUMBridge v0.1 中的桥接亚型分类。

ENTITY-MERONOMY 这个回指短语是先行词的一个子部分（或反之），即回指短语和先行词之间存在某种部分-整体关系。

我在湖边看到了 a large house。门是红色的。

ENTITY-PROPERTY 这个指示词是前件的一个物理或无形的属性（或者反过来）。例如：气味、长度、风格等。

我拿起了 a bouquet of roses。那个气味很香。

ENTITY-RESULTATIVE 回指项可以从先行项逻辑推断出来（或反之亦然）。这通常是由变换或产品生成过程引起的结果，例如烹饪。²

虽然 my flour 是一个奇怪的质地，但面包出来得很完美。

SET-MEMBER 回指是前件集合的一个元素（或反之亦然）。

我在生日时得到了 several books。这本悬疑小说是我的最爱。

回指词是前件集的一个子集（或反之亦然）。

(1) A group of students 进入了大厅。男孩子们穿着制服，系着领带。

SET-SPAN-INTERVAL 指代词是空间或时间先行词区间的一个子跨度（或反之亦然）。

如果你想在 Sunday 见面，我上午有空。

OTHER 指示语和先行语符合识别桥接对的标准，但不属于上述任何桥接子类型。例如，Ogrodniczuk and Zawislawska (2016) 提供了元引用的例子：

²这种子类型包含了 Fang et al. (2022) 专门为食谱结果提出的 TRANSFORMED 类型。

我昨天去了 Sensational Cakes，但我觉得蛋糕不是很好吃。

元指涉允许对名称或标签进行回指，如例子 (1)。这类实例既独特又有趣，不希望将其勉强归入另一个类别，但也不够常见，无法在子类型分类中设立单独的类别。

如 3.1 节所述，识别桥接实例的标准是指示性的，依赖于信息状态并回溯到关联的先行词。子类型标签主要帮助我们理解这种现象在话语中的表现，因此，理论上没有理由将适用于桥接实例的子类型限制为仅一个。实际上，有些桥接的情况可能会同时适用于多个子类型：

(2) Several women 走进了房间。一个人立刻离开了。

我会来拜访 this week，因为我上周没能来。

示例 (2) 显示了一个实例，其中 COMPARISON-SENSE 和 SET-MEMBER 都适用，而示例 (2) 显示了一个 COMPARISON-RELATIVE 和 COMPARISON-TIME 适用的情况。在这个标注试点中，标注者被指示选择一个单一的连接子类型，并在多个类别同时出现时优先选择某些类别。然而，原则上，所有适用的子类型都可以被标注。在我们随后的努力中，我们打算在 GUM 的剩余数据上进行标注并推出完整版本的 GUMBridge，我们计划支持对整个语料库中单个连接对的多个连接子类型的标注。

3.3 注释过程

GUMBridge 标注试点是在现有的 GUM (v10) 语料库的测试集上进行的，该语料库由 26 篇文档 (~26k tokens) 组成，涵盖 16 个体裁（学

术写作、传记、法庭记录、散文、小说、指南、采访、书信、新闻、在线论坛讨论、播客、政治演讲、临时面对面谈话、教科书、旅游指南和视频博客)。GUM 语料库已经包括实体范围、共指、³ 和信息状态的标注, 即“新”、“给定”和“可接触”(不包括桥接实例的可接触性)。

测试集的文档都进行了双重标注, 本论文的一位作者作为标注者 A, 而不同的语言学研究作为标注者 B 为测试集中的不同文档标注。担任标注者 B 的 8 位标注者中, 每位被分配到 2 到 4 份测试集文档。标注是使用 GitDox 标注界面 (Zhang and Zeldes, 2017) 完成的。对于文档中现有的实体标注, 标注者被要求识别实体是否为桥接照应, 如果是, 则需要在照应词和其关联前述体之间创建链接。标注者还需要将桥接照应词的 IS 更新为“可获取”并选择照应词的桥接子类型标注。提供给标注者的完整标注指南包含在补充材料中。

3.4 一致性研究

在表 2 中, 我们提供了桥接注释过程三个阶段的一致性数值: 指代词识别、先行词解析和子类型分类。

	Precision	Recall	F1 Score
Anaphor Recognition	0.44	0.34	0.38
Anaphor+Antecedent Recognition	0.32	0.25	0.28
Accuracy			
Antecedent Resolution		0.72	
Cohen's κ			
Bridging Subtype		0.58	

Table 2: GUMBridge 初步标注者间一致性

对于搭桥对(指代词 + 先行词)的识别以及仅识别搭桥指代词, 我们给出了标注员 B 相对于标注员 A 的 PRF。我们看到搭桥指代词识别的 F1 为 0.38, 而搭桥对识别的 F1 仅为 0.28。由于搭桥对的识别本质上受限于指代词的识别, 我们还给出了当两个标注员在搭桥指代词上达成一致时, 标注员 B 选择先行实体的准确率, 总共 133 个案例中有 72 个是正确的。最后, 在两个标注员均就搭桥对的指代词和先行词达成一致的 96 个实例中, 搭桥子类型标注的 Cohen's Kappa 为 0.58, 表明中等一致性。这些数据表明关键的障碍在于指代词识别, 尽管先行词解析和子类型标注也不是简单的问题。

³共指方案考虑了所有符合桥接条件的提及, 包括不定指代词、非名词性前指的语篇指示等, 详细讨论见 Zeldes (2022)。

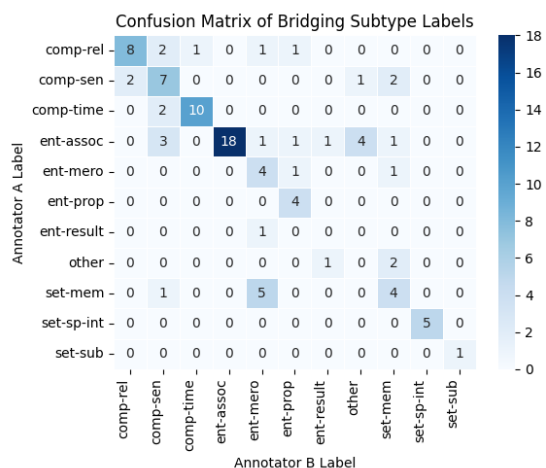


Figure 2: 在搭桥实例中, 通过混淆矩阵进行子类型区分, 其中先行词和指代词的注释匹配。

在图 2 中, 我们展示了由标注者 A 和标注者 B 分配给重叠桥接对的桥接子类型标签的混淆矩阵。我们看到, 重叠最多的子类型是 COMPARISON 类别和 ENTITY-ASSOCIATIVE。虽然有些类别存在意见分歧, 并且这些分歧涵盖了多个类别, 我们注意到 ENTITY-MERONOMY 和 SET-MEMBER 类别特别容易混淆, 这表明部分-整体关系和集合-成员关系可能非常相似。ENTITY-ASSOCIATIVE 和 OTHER 类别也特别容易混淆, 这说明 ENTITY-ASSOCIATIVE 可能是一个过于宽泛的类别。尽管桥接子类型标注上的一致性是中等的, 但显然对于这些类别的指导原则仍需进行细化。然而, 由于桥接实例识别上的一致性明显较低, 桥接回指识别成为标注过程中的限制点。

3.5 数据评判

如前节所示, 标注试点的结果显示标注者之间的一致性较低, 因此需要对标注进行定性分析以确定分歧原因。作为此过程的一部分, 试点中的标注经过裁定, 生成了用于 GUMBridge 测试集 (v0.1) 的单一参考桥接标注集, 该标注集根据 Creative Commons Attribution (CC-BY) 4.0 版许可协议随本文发布。裁定后按桥接子类型组成的 GUMBridge 测试集在附录 A 中显示。GUMBridge 测试集中共有 401 个桥接标注, 平均每千个标记有 15.4 个桥接实例。这与 ISNotes 和 ARRAU RST 中每千个标记中找到的更高桥接实例率相当, 如表 1 所示。尽管在该试点中标注的数据集规模有限限制了我们对其效应进行观察, 但为了完整性, 在附录 B 中包括了每种体裁中观察到的桥接关系类型的细分。

值得注意的是, GUM (v10) 标注的测试集

Completely Matching	61
Different Subtype	35
Different Antecedent	37
Annotator B Only	172
Annotator A Only	257
Total	562

Table 3: GUMBridge 初步注释中标注者一致/不一致类型的计数。

中实例数量几乎翻了一倍，从 222 个搭桥实例增加到 GUMBridge 测试中的 401 个，表明在新的标注工作中对搭桥实例的覆盖有了显著提高。尽管与 2 部分讨论的一些标注工作相比，这次标注的连贯性较差，但数据表明召回率更高，这使我们能够捕捉到更广泛的搭桥实例范围。由于搭桥现象通常较为稀少，即使初始一致性较低，这些标注也可以在裁定过程中进行人工审核和验证。因此，我们认为，倾向于采用高召回率的标注方法并在审核时消除误判比冒着许多有趣案例无法识别的风险要优越。

判定过程包括在指示代词、前指和子类型层面比较标注者 A 和标注者 B 的所有分歧判断。表 3 显示了在试点标注中此类分歧的比例。在标注者 B 标记为桥接而标注者 A 初始根本未标记为桥接的 172 个实例中，经过重新评估后，结果认为其中 64 个 (37%) 可以合理地被视为桥接形式。这些判断中的许多依赖于对涉及话语实体的主观理解。在接下来的部分中，我们将分析主观性在此标注试点中的影响，以及如何在未来更好地处理这一问题。

4 桥接标注中的主观性

在语言数据开发过程中关于主观性的先前研究，通常集中在标注者判断可能有很高变异性的领域，例如仇恨言论检测和情感分析（例如，Waseem (2016); Kenyon-Dean et al. (2018)），尽管一些研究也关注于那些看似更加客观的任务，如词性标注（例如，Plank et al. (2014)）。一些研究讨论了在标注工作中包含主观判断的范式及其影响，而不是试图消除所有模糊性（Ovesdotter Alm, 2011; Röttger et al., 2022）。最终，合适的方法取决于手头的语言任务以及研究人员希望通过标注工作达到的目标。

尽管本文的标注试点为标注人员提供了详细的指导原则，但主观判断仍然是桥接实例标注的内在部分，因为标注人员在做决定时依赖于他们对话语中实体之间隐含关系的理解。正如前面所指出的，在桥接实例标注中有三个决策点可能引入主观判断：(1) 识别桥接照应词，(2) 确定相应的关联前件，(3) 选择该对的桥接子类型类别。以下各节提供示例，以说明每个标

注阶段中存在的关于主观性的独特考虑。

4.1 子类型分类

选择桥接子类型类别依赖于理解桥接配对中回指词和先行词之间的关系。两个实体之间关系的确切性质取决于标注者对这两个实体的主观理解。对相关实体的不熟悉可能导致标注错误：

the cuttings → 第一个垫子

在例子 4.1 中，“切割”指的是仙人掌切割，每个切割都是一个完整的仙人掌叶片。在没有这项具体知识的情况下，标注者可能会合理地假设一个叶片是切割的一部分或者一个切割是叶片的一部分。

在根据话语的上下文来解释实体时，可能会有额外的不确定性：

peppermint plants → 薄荷

在例子 4.1 的语境下，不清楚“薄荷”是指回薄荷植物的特定部分（例如叶子），还是一种借代的方式，指代整个植物。

在语境中也存在可能出现多种子类型的情况：

(3) some basil → 种子

在例子 (3) 的语境中，提出了一个问题，即是否可以从“种子”中生长出“一些罗勒”。因此，可以合理地说罗勒来自种子，在这种情况下，子类型将是 ENTITY-RESULTATIVE。然而，也可以合理地说种子是罗勒植物的一部分，在这种情况下，子类型将是 ENTITY-MERONOMY。在这种情况下，有必要有一个优先级层次来决定应该分配哪个桥接子类型类别，或者我们必须允许多个子类型标注。在未来的工作中，我们打算为整个 GUMBridge 语料库支持多个桥接子类型的标注。

4.2 前件选择

当标注者选择桥接照应的关联前件时，也有机会进行主观判断。在某些情况下，多个先行的实体可能是桥接前件的合理候选者：

your mouth → 其他身体部位…

teeth → 其他身体部位…

例子 4.2 指的是制作牙模的情形，叙述者想知道身体的其他部位是否可以接受相同的处理。尚不清楚“其他身体部位”是更适合与“嘴”对比，还是与“牙齿”对比，或者甚至是两者皆可接受为身体部位。

也存在对指代词指称的分歧的可能性：

the bridge → 边缘

the upper levels → 边缘

在例子 4.2 中，叙述者考虑“越过边缘”，但不清楚这是否是指某个特定桥梁的边缘，或者它是指某个一般的高层边缘。在这种情况下，可能有利于施加一个易于执行的启发式方法，例如选择离桥接照应词更近的选项，假设我们的目标是进行一次单一的指代决策。注意，这与多个标签适用的情况不同，因为这两种解释虽然都可能，但却是互斥的。

4.3 指示语识别

在识别桥接照应时，标注者必须主观判断一个实体是否因世界知识而可访问（因此不构成桥接）或该可访问性是否可归因于前面提到的一个实体。例如，一位标注者将“列西普和德漠克里特”从“古希腊哲学家”中桥接，但不对更广为人知的“亚里士多德”进行桥接。这表明标注者的世界知识可能以一种不希望的方式影响他们对“可访问”的看法，因为这会导致标注者之间的不一致。我们建议，泛指/世界知识的可访问性应与知识库（如维基百科）相关联，而不是由各个标注者自行判断。对于命名实体，这种链接或维基化已经在 GUM (Lin and Zeldes, 2021) 中可用，并将在未来的标注工作中整合。

5 结论

在本文中，我们研究了标注者判断中的主观性对标注桥接实例各个阶段的影响。我们利用试点项目产生的标注来创建一个新的桥接标注资源，GUMBridge，以进行研究。我们还发布了 GUMBridge (v0.1) 初步测试集的桥接标注裁定版本。在后续工作中，我们计划完善试点中使用的指南和标注程序，随后将用于标注 GUM 语料库的剩余部分（开发和训练），以生成 GUMBridge 的完整版本，并扩展我们的标注到 GUM 的超领域挑战测试集 GENTLE（语言评估的体裁测试，Aoyama et al. 2023）。由于手动标注桥接所需的时间和精力限制了标注过程的可扩展性，我们还将探讨结合半自动化方法，例如结合 LLMs 或其他桥接解决系统与人工修正，以提高过程效率。

在我们开发 GUMBridge 测试 (v0.1) 期间，我们发现注释者在选择桥接对的子类型方面的协议程度适中，但在识别桥接指代方面更难达成一致。这表明，桥接指代的识别是注释过程中最容易受到注释者主观判断影响的阶段，因此在考虑注释者的主观性时应给予最大的关注。虽然一些主观性源于语境中的语言固有的模糊性，但其他方面的主观性可以通过提供指导方针来解决，以便在有多种选择时决定优先判断。

局限性

本文中关于桥接照应注释中的主观性的分析基于尚在开发中的一项新资源的初步注释研究。这限制了可用于分析的数据量，仅限于一个包含 26k 标记的测试集。注释方案的可靠性也是一个限制，因为注释试验的结果显示在桥接照应识别上的一致性不理想，并且在未来的工作中需要对注释方案/说明进行修订。

References

- Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. [GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation](#). In Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 166–178, Toronto, Canada. Association for Computational Linguistics.
- Herbert H. Clark. 1975. [Bridging](#). In Theoretical Issues in Natural Language Processing.
- Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. [A Discourse Information Radio News Database for Linguistic Analysis](#), pages 65–76. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. [What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text](#). In Findings of the Association for Computational Linguistics: ACL 2022, pages 3481–3495, Dublin, Ireland. Association for Computational Linguistics.
- Yulia Grishina. 2016. [Experiments on bridging across languages and genres](#). In Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), pages 7–15, San Diego, California. Association for Computational Linguistics.
- Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhandari, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. [Sentiment analysis: It's complicated!](#) In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.
- Hideo Kobayashi, Yufang Hou, and Vincent Ng. 2023. [PairSpanBERT: An enhanced language model for bridging resolution](#). In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6931–6946, Toronto, Canada. Association for Computational Linguistics.

- Hideo Kobayashi and Vincent Ng. 2020. [Bridging resolution: A survey of the state of the art](#). In Proceedings of the 28th International Conference on Computational Linguistics , pages 3708–3721, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Lauren Levine and Amir Zeldes. 2024. [Unifying the scope of bridging anaphora types in English: Bridging annotations in ARRAU and GUM](#). In Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference , pages 41–51, Miami. Association for Computational Linguistics.
- Jessica Lin and Amir Zeldes. 2021. [WikiGUM: Exhaustive entity linking for wikification in 12 genres](#). In Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop , pages 170–175, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Katja Markert, Yufang Hou, and Michael Strube. 2012. [Collective classification for fine-grained information status](#). In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.
- Natalia N Modjeska. 2004. Resolving other-anaphora.
- Anna Nedoluzhko, Jiří Mírovský, and Petr Pajas. 2009. [The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague dependency treebank](#). In Proceedings of the Third Linguistic Annotation Workshop (LAW III) , pages 108–111, Suntec, Singapore. Association for Computational Linguistics.
- Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. [An annotation scheme for information status in dialogue](#). In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04) , Lisbon, Portugal. European Language Resources Association (ELRA).
- Tim O’Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. [Richer event description: Integrating event coreference with temporal, causal and bridging annotation](#). In Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016) , pages 47–56, Austin, Texas. Association for Computational Linguistics.
- Maciej Ogrodniczuk and Magdalena Zawisławska. 2016. [Bridging relations in Polish: Adaptation of existing typologies](#). In Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016) , pages 16–22, San Diego, California. Association for Computational Linguistics.
- Cecilia Ovesdotter Alm. 2011. [Subjective natural language problems: Motivations, applications, characterizations, and implications](#). In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies , pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. [Linguistically debatable or just plain wrong?](#) In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) , pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.
- Massimo Poesio. 2004. [Discourse annotation and semantic annotation in the GNOME corpus](#). In Proceedings of the Workshop on Discourse Annotation , pages 72–79, Barcelona, Spain. Association for Computational Linguistics.
- Massimo Poesio and Ron Artstein. 2008. [Anaphoric annotation in the ARRAU corpus](#). In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08) , Marrakech, Morocco. European Language Resources Association (ELRA).
- Ellen F. Prince. 1981. Toward a taxonomy of given-new information. *Radical pragmatics* , pages 223–255.
- Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2010. [A typology of near-identity relations for coreference \(NIDENT\)](#). In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) , Valletta, Malta. European Language Resources Association (ELRA).
- Ina Roesiger. 2016. [SciCorp: A corpus of English scientific articles annotated for information status analysis](#). In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) , pages 1743–1749, Portorož, Slovenia. European Language Resources Association (ELRA).
- Ina Rösiger. 2018. [BASHI: A corpus of Wall Street Journal articles annotated with bridging links](#). In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) , Miyazaki, Japan. European Language Resources Association (ELRA).
- Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien, and Jonas Kuhn.

2018. [German radio interviews: The GRAIN release of the SFB732 silver standard collection](#). In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) , Miyazaki, Japan. European Language Resources Association (ELRA).

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Joseba Rodríguez, and Massimo Poesio. 2019. [Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus](#). Natural Language Engineering , 26:95 – 128.

Zeera Waseem. 2016. [Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter](#). In Proceedings of the First Workshop on NLP and Computational Social Science , pages 138–142, Austin, Texas. Association for Computational Linguistics.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium , 17.

Juntao Yu, Sopan Khosla, Ramesh Manuvinaurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2022. [The CODI-CRAC 2022 shared task on anaphora, bridging, and discourse deixis in dialogue](#). In Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue , pages 1–14, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Amir Zeldes. 2017. [The GUM corpus: Creating multilayer resources in the classroom](#). Language Resources and Evaluation , 51(3):581–612.

Amir Zeldes. 2022. [Can we fix the scope for coreference? Problems and solutions for benchmarks beyond OntoNotes](#). Dialogue & Discourse , 13(1):41–62.

Shuo Zhang and Amir Zeldes. 2017. [GitDOX: A linked version controlled online XML editor for manuscript transcription](#). In Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2017) , pages 619–623.

A GUMBridge 测试中的子类型

表 4 显示了 GUMBridge 测试 v0.1 裁定版本中桥接子类型的计数。

B GUMBridge 测试中的体裁亚型

图 3 显示了在 GUMBridge 测试集 (v0.1) 的 16 个体裁中，每 1k 词标的每种桥接关系类型 (COMPARISON、SET、ENTITY 和 OTHER) 的桥接实例数量。

COMPARISON	
RELATIVE	59
TIME	27
SENSE	45
Subtotal	131
ENTITY	
ASSOCIATIVE	124
MERONOMY	37
PROPERTY	9
RESULTATIVE	21
Subtotal	191
SET	
MEMBER	31
SUBSET	14
SPAN-INTERVAL	18
Subtotal	63
OTHER	16
Total	401

Table 4: 在裁定的 GUMBridge 数据中桥接子类型的计数。

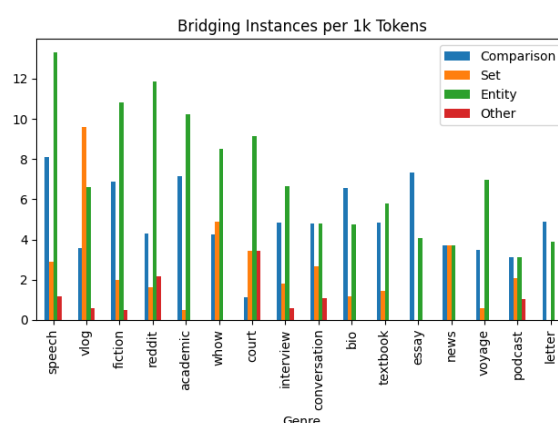


Figure 3: 在已判定的 GUMBridge 数据中按体裁划分的桥接关系类型的计数。

C 与 ARRAU 桥接子类型的比较

为了更好地比较 GUMBridge 和 ARRAU 的资源，我们简要比较了 ARRAU 的桥接子类型⁴如何映射到 GUMBridge 的提议模式：

大多数属于 ENTITY-MERONOMY 或 ENTITY-PROPERTY 的部分关系。

元素集 → 映射到 SET-MEMBER 。

子集集合 → 映射到 SET-SUBSET 。

‘其他’照应 → 映射到 COMPARISON-RELATIVE，其中包括了 ARRAU 中未涵盖的额外比较标记，包括序数和比较级形容词。

未详细指定的 → ENTITY-ASSOCIATIVE，除非根据上下文，有其他 ENTITY 子类型更适合。然而，感觉照应 (green shirt → 红色的那个) 应该被映射到 COMPARATIVE-SENSE 。

⁴由于 GUMBridge 框架在子类型关系中不区分回指词和先行词的相对角色，所以 ARRAU 的反向子类型与其常规子类型的映射相同。