众包隐式话语关系的设计选择：揭示任务设计引入的偏差

Valentina Pyatkin[1] FrancesYung[2] Merel C.J.Scholman[2,3]
ReutTsarfaty[1] Ido Dagan[1] Vera Demberg[2][1]以色列拉马特甘，巴伊兰大学[2]
德国萨尔布吕肯，萨尔兰大学[3]荷兰乌得勒支，乌得勒支大学
{pyatkiv,reut.tsarfaty}@biu.ac.il; dagan@cs.biu.ac.il
{m.c.j.scholman,frances,vera}@coli.uni-saarland.de

arXiv:2304.00815v1 [cs.CL] 3 Apr 2023

摘要

自然语言标注中的分歧，大多是从标注者和标注框架所引入的偏差这一角度进行研究的。本文提出分析另一种偏差来源：任务设计偏差。这种偏差对众包语言学标注的影响尤为显著，因为这类标注使用自然语言来引导非专业标注者进行解读。为此，我们考察隐式话语关系标注；由于关系存在歧义，这项任务已被反复证明颇具难度。我们比较了采用两种不同标注任务获得的 1,200 条话语关系标注，并量化了这两种方法在四个不同领域中的偏差。这两种方法都是面向众包设计的自然语言标注任务。我们表明，任务设计会使标注者倾向于某些关系，而且某些话语关系类型可以通过其中一种标注方法比另一种更好地引出。我们还得出结论，在模型训练和测试时应将此类偏差纳入考量。

1 引言

众包已成为一种流行的数据收集方法。它不仅能让研究人员在更短时间内收集大量标注数据，还能捕捉人类对自然语言的推理，而这应当是基准自然语言处理任务的目标（Manning，2006年）。为了获得可靠标注，传统上会将众包标签汇总为每个项目的单一标签，采用简单的多数投票，或根据标注者之间的分歧使用标注模型来降低数据中的噪声（Hovy 等人，2013；Passonneau 和 Carpenter，2014年）。然而，越来越多的人达成共识：在各种自然语言处理任务中，标注分歧通常不能被视为噪声并予以舍弃，例如自然语言推理（De Marneffe 等人，2012；Pavlick and Kwiatkowski，2019；Chen 等人，2020；Nie 等人，2020）、词义消歧（Jurgens，2013）、问答（Min 等人，2020；Ferracane 等人，2021）、照应消解（Poesio 和 Artstein，2005年；Poesio 等人，2006年）、情感分析（Díaz 等人，2018年；Cowen 等人，2019年）以及立场分类（Waseem，2016年；Luo 等人，2020年）。有人提出用标签分布取代类别标签，以表示标签歧义（Aroyo 和 Welty，2013；Pavlick and Kwiatkowski，2019；Uma 等人，2021年；Dumitrache 等人，2021年）。

语言标注存在歧义的背后有多种原因（Dumitrache，2015；Jiang and de Marneffe，2022年）。Aroyo 和 Welty（2013）将歧义来源概括为三类：文本、标注者和标注方案。在下游自然语言处理任务中，如果模型能够检测歧义文本可能存在的不同解释，或预测某一群体的解释分布，将会很有帮助。除了现有关于标注者偏差导致的分歧的研究之外，标注框架的影响也得到了研究，例如关于蕴含是否应包括语用推理的讨论（Pavlick andKwiatkowski，2019年）、所收集标签的粒度所产生的影响（Chung 等人，2019年），或用于对语言现象进行分类的标签体系（Demberg 等人，2019年）。在本研究中，我们考察了任务设计偏差对众包标注质量的影响；这种偏差独立于标注框架。具体而言，我们关注句间隐式话语关系（话语关系）标注，即两个相邻句子之间不存在可将关系意义归因于其上的话语连接词时，二者之间的语义或语用关系。图1展示了一个可标注为

合取还是结果。

隐式话语关系标注可以说是话语解析中最困难的任务。话语连贯性是读者对文本形成的心理表征的一项特征，而不是语言材料本身的特征（Sanders 等人，1992）。因此，话语标注依赖于标注者对文本的解释。此外，关系往往可以有多种解释（Rohde 等人，2016年），并且多种有效解读可以同时成立。这些因素使话语关系标注，尤其是隐式关系标注，成为一项特别困难的任务。对于每个话语关系，我们收集了10份不同的标注，从而聚焦于分布式表征；与类别标签相比，分布式表征包含的信息更丰富。

由于话语关系标注中的标签往往是抽象术语，非专业人士不易理解，因此我们聚焦于“自然语言”任务设计。研究表明，将标注任务拆解并简化，使话语关系标签能够通过自然语言标注间接获得，有助于开展众包（Chang 等人，2016年；Scholman 和 Demberg，2017年；Pyatkin 等人，2020年）。采用自然语言的众包日益普及，包括自然语言推理（Bowman 等人，2015）、语义角色标注（Fitzgerald 等人，2018）和问答（Rajpurkar 等人，2018）等任务。这一趋势也体现在建模方法中：这些方法将传统的结构化预测任务转化为自然语言任务，例如共指消解（Aralikatte 等人，2021年）、话语理解（Ko 等人，2021年）或桥接回指（Hou，2020年；Elazar 等人，2022年）。因此，更广泛的研究界有必要了解，即使任务对非专业人士更易上手，任务设计偏差仍可能如何产生。

我们考察了两项不同的自然语言众包话语关系标注任务（图 1）：Yung 等人（2019年）根据众包工作者插入的话语连接词（直接分类）推导关系标签；Pyatkin 等人（2020年）则根据众包工作者撰写的问答（QA）对推导标签。两种任务设计都采用自然语言标注，而非分类体系中的标签。两种任务设计——直接分类和问答——用于标注4个不同领域中的1,200个隐式话语关系。这使我们能够探究任务设计如何影响所得标注，以及每种方法固有的偏差。为此，我们展示了各种

标注上的标注者间一致性指标

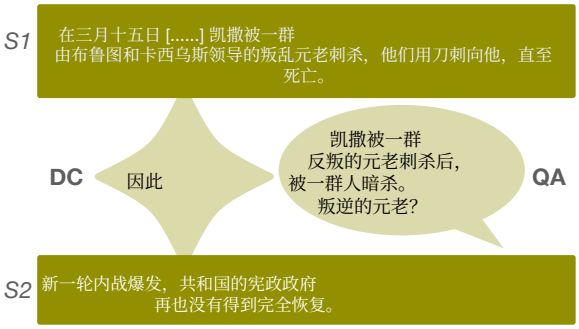


图1：两个关系论元（S1 和 S2）以及中间的直接分类（DC）和问答（QA）标注示例。

采用分布式标签和聚合标签。

我们发现，两种方法在识别某些类型的关系时各有优缺点。我们进一步发现，这些偏差也会受到领域的影响。在一系列话语关系分类实验中，我们展示了采用混合方法收集标注的优势；我们表明，以分布作为目标、使用软损失进行培训能够提升模型性能；并且我们发现，跨任务泛化比跨领域泛化更难。

本文的结构如下。我们介绍任务设计偏差这一概念，并使用两种不同的任务设计分析其对众包隐式话语关系的影响（第3-4节）。接下来，我们利用获得的标注量化每种方法的优势与不足，并提出减少任务偏差的方法（第5节）。随后，我们考察体裁特定任务偏差（第6节）。最后，我们展示任务偏差对话语关系分类性能的影响（第7节）。

2 背景

2.1 标注偏差

标注往往是一项本质上具有歧义的任务，通常存在多种可能的解释，且没有唯一的真实标注（Aroyo 和 Welty，2013）。越来越多的研究关注标注分歧和偏差。

以往研究关注众包工作者可能受到哪些偏见影响。工作者偏见受多种因素影响，例如教育或文化背景，以及其他人口统计特征。Prabhakaran 等人(2021年) 指出，对于更具主观性的标注任务，标注者的社会人口背景会带来多种标注视角，并

认为标签聚合会掩盖这些视角。因此，有人提出使用软标签，例如 CrowdTruth 方法提供的软标签（Dumitrache 等人，2018年），这种方法要求为每个实例收集多个判断（Uma 等人，2021年）。Bowman 和 Dahl (2021年) 建议，不应将受方法学伪影偏差影响的标注纳入基准数据集。相比之下，Basile 等人(2021年) 认为，自然语言理解模型应预测各种形式的人类分歧，因此这些分歧应纳入评测数据集。

与标注者偏差相比，关于任务表述所引发偏差的研究较少。Jakobsen 等人(2022年)表明，论证标注所呈现的社会群体差异程度因标注者遵循的指南不同而大相径庭。类似地，Buechel 和 Hahn(2017a,b)研究了众包情绪标注中的不同设计选择，并表明指南要求标注者采取的视角会影响标注质量和分布。Jiang 等人(2017年)研究了释义收集工作流程的影响，发现基于先前贡献的示例会促使标注人员产出差异更大的释义。Hube 等人(2019年)表明，让标注人员思考其他标注人员可能给出的回答，并使他们意识到自身可能存在的偏差，可以减轻主观判断标注中的偏差。因此，现有研究表明，任务设计会以多种方式影响标注结果。后续研究考察了多个标签的收集：Jurgens(2013)比较了选择与量表评分，发现标注人员在词义标注任务中会选择额外的标签。相比之下，Scholman 和 Demberg (2017年) 发现，即使允许标注人员提供额外的话语关系标签，他们通常也不会这样做。Chung 等人 (2019年)比较了多种标签收集方法，包括单标签／多标签标注、排序和概率分配。我们关注的是采用同一组标签、但将其译成不同“自然语言”以用于众包的话语关系标注方法中的偏差。

2.2 话语关系标注

有多种框架可用于标注话语关系，例如修辞结构理论（Mann 和 Thompson，1988年）和分段话语表示理论（Asher，1993）。

在这项工作中，我们聚焦于隐式话语关系的标注，并采用用于标注宾夕法尼亚话语树库 3.0（PDTB，Webber 等人，2019年）的框架。PDTB 的关系类型分类采用三级层级结构：第一级包含四个粗粒度的关系类型组，后续每一级则包含更细粒度的关系类型。[1] 这一流程结合了人工标注和自动标注：自动流程会识别潜在的显式连接词，随后由标注者判断这些潜在连接词是否确实为真实连接词。如果是，他们会指定其论元之间存在的一种或多种关系类型。如果不存在连接词或其他词汇表达形式（即对于隐式关系），每位标注者都会提供一个或多个连接词，以共同表达他们推断出的关系类型。

话语关系数据集，如PDTB（Webber等人，2019）、RST-DT（Carlson和Marcu，2001年）和TED-MDB（Zeyrek等人，2019），通常由受训标注者进行标注；这些标注者需要熟悉针对特定任务编写的详尽指南（Plank等人，2014；Artstein和Poesio，2008年；Riezler，2014年）。不过，也有研究尝试通过众包进行话语关系标注（Kawahara等人，2014；Kishimoto等人，2018；Scholman和Demberg，2017；Pyatkin等人，2020）。我们研究了两种对句间隐式话语关系进行标注的众包方法，并将自然语言标注以确定性方式映射到PDTB3标签框架。

2.2.1 使用DC方法众包标注话语关系

Yung 等人（2019年）开发了一种利用话语连接词进行话语关系标注的众包方法，称为 DC 方法。对于每个实例，参与者首先提供一个他们认为最能表达两个论元之间关系的连接词。需要注意的是，参与者选择的连接词可能存在歧义。因此，参与者会在第二步消除关系歧义：从根据第一步所提供的连接词动态生成的列表中选择一个连接词。如果第一步填入的内容与连接词库中的任何条目都不匹配（该连接词库用于生成消歧连接词列表），系统就会向参与者提供一个默认列表，其中包含十二个表达各种关系的连接

[1]我们将信念与言语行为关系义（无法通过问答和直接分类可靠地区分）与其对应的更一般关系义合并。

词。根据两个步骤中选择的连接词，可以提取推断出的关系义项。例如，图1 中的合取解读可用此外来表达，而结果解读可用因此来表达。

DC方法被用于创建一个包含6,505个经话语关系标注的隐含关系的众包语料库，名为DiscoGeM（Scholman 等人，2022a）。本研究使用了 DiscoGeM 的一个子集（见第3节）。

2.2.2 通过问答方法对话语关系进行众包

Pyatkin 等人 (2020年) 提出使用问答对对话语关系进行众包。他们收集了一组句内问答标注数据，旨在通过在问题中包含一个命题、在相应答案中包含另一个命题来表示话语关系；问题前缀（什么与……相似？什么是……的一个例子？）则对应一种关系义项。后来，他们将该方法扩展至句间场景（Scholman 等人，2022b）。本文采用了扩展后的方法，通过一个问题和答案关联两个不同的句子。例如，以下问答对将图 1中的两个句子联系起来，构成一种结果关系。

(1) **被** 凯撒遭一群叛乱参议员暗杀的结果是什么？(S1) - 一系列新的内战爆发了 *[...]*(S2)

标注流程包括以下步骤：给定两个相邻句子，标注者需要选择一个句子，用来构造一个问题。另一个句子则作为该问题的答案。接下来，他们通过选择问题前缀，并从选定的句子中提取内容来补全问题，从而构造出一个问题。

由于对于一组特定的对称关系，两个句子都可以被选作问题或答案（即 **一系列新的内战爆发的原因** ），我们将这两种可能的表述视为等同。

可能的问题前缀涵盖了 PDTB 3.0 中的所有关系类型（不包括信念关系和言语行为关系）。关系义项的方向，例如论元*1*作为否定项 与 论元*2*作为否定项，取决于哪一个句子被选作问题或答案。虽然Pyatkin 等人（2020年）允许众包工作者为每个实例构建多个问答对，即为每个关系标注多个话语关系类型，但我们决定将任务限制为每位标注人员对每个关系只标注 1 种关系类型。我们做出这一决定，是为了让问答方法与直接分类方法更具可比性；直接分类方法也只允许插入一个连接词。

3 方法

3.1 数据

我们使用直接分类和问答任务设计，对 1,200 个句间话语关系进行了标注。[2]这 1,200 个关系中，900 个取自 DiscoGeM 语料库，300 个取自 PDTB 3.0。

DiscoGeM 关系 本研究纳入的 900 个 DiscoGeM 实例来自不同领域：其中 296 个实例取自 DiscoGeM 关系子集，该子集来源于 Europarl 会议记录（取自 Europarl 语料库的政治演讲准备稿的书面记录；Koehn，2005年），304 个实例取自文学子集（五本英文书籍中的叙事文本），[3] 另有 300 个实例取自 DiscoGeM 的维基百科子集（信息性文本，选自 30 篇维基百科文章的摘要）。这些不同体裁使跨体裁比较成为可能。鉴于某些关系类型的出现频率可能因体裁而异，这种比较是必要的（Rehbein 等人，2016年；Scholman 等人，2022a；Webber，2009年）。

这900个关系已在DiscoGeM中使用直接分类方法完成标注；本研究还使用问答方法收集标签。除了通过直接分类和问答方法收集众包标签外，维基百科子集还由三名受训标注者进行标注。[4]这些维基百科实例中，47%被专家标注者赋予了多个关系义项（即被认为存在歧义或表达了多种解读）。

[2]标注可在以下网址获取：https://github.com/merelscholman/DiscoGeM。[3]《动物农场》，作者乔治·奥威尔；《哈利·波特与魔法石》，作者 *J. K.* 罗琳；《银河系漫游指南》，作者道格拉斯·亚当斯；《了不起的盖茨比》，作者 *F*·斯科特·菲茨杰拉德；以及《霍比特人》，作者 J. R. R. 托尔金[4]实例由两名标注者进行标注，并由第三人核验；Cohen's κ 第一名标注者与参考标签之间的一致性为 .82（一致率为 88%），第二名标注者与参考标签之间的一致性为 .96（一致率为 97%）。更多详情请参见Scholman 等人(2022a)。

PDTB关系 纳入PDTB关系的目的是将我们的标注与传统的PDTB金标准标注进行比较。这些实例（均为句间实例）经过筛选，覆盖所有关系类别；每类随机抽取最多15个、至少2个关系实例（对于关系实例少于15个的类别，则抽取所有现有关系）。PDTB实例的参考标签由作为PDTB3语料库一部分标注的原始PDTB标签构成。其中只有8%包含多个关系义项。

3.2 众包工作者

通过 Prolific 采用筛选方式招募众包工作者（Scholman 等人，2022b）。研究表明，这种方式能在 DR 标注的质量与时间／金钱成本之间取得良好平衡。众包工作者须满足以下要求：以英语为母语，居住在英国、爱尔兰、美国或加拿大，并至少取得本科学位。

符合这些条件的标注人员可以参加初始招募任务，他们需要使用 DC 或 QA 方法之一标注一段文本，并会立即收到表现反馈。在该任务中准确率达到 ≥ 0.5 的标注人员，才有资格参加后续任务。因此，我们为每种方法分别建立了一组独立的众包工作者。DC 标注（作为 DiscoGeM 项目的一部分收集）最终由 199 名入选众包工作者提供；QA 则由 43 名入选众包工作者完成。[5] 生产数据收集期间，我们持续监控质量，并根据表现调整资格要求。

每个方法下的每个实例均由 10 名标注人员进行标注。这个人数是为了与以往研究保持一致而选定的。例如，Snow 等人(2008年) 表明，对于多种语言学标注任务，每个实例采用 10 份众包标注即可获得令人满意的准确率。Scholman 和 Demberg (2017年) 发现，指派一组新的 10 名标注者来标注相同实例，几乎完美地复现了早期一项 DC 研究中的连接词插入结果。

实例按每批 20 个进行标注。对于问答，每批大约需要 20 分钟完成，而直接分类每批需要 7 分钟。标注人员每批分别获得约 2.50 英镑和 1.88 英镑的报酬。

3.3 标注者间一致性

我们通过两种话语关系标注方法所收集标注之间的标注者间一致性（IAA），以及这些标注与受训标注者所收集参考标注之间的一致性，来评估这两种话语关系标注方法。

Cohen's κ系数（Cohen，1960）是一种常用于衡量标注者间一致性（IAA）的指标。对于话语关系标注，Cohen's κ系数达到 .7 通常被认为体现了良好的标注者间一致性（Spooren 和 Degand，2010年）。然而，以往研究表明，相较于显式关系，隐式关系更难达成一致：Kishi- moto 等人(2018年) 报告称，使用包含 7 个二级标签的标签集对众包隐式关系标注进行评估，F1 值为 .51；Zikánová 等人 (2019年) 报告称 κ=，使用包含 23 个二级标签的标签集对隐式关系进行专家标注，所得结果为 .47（58%）；而Demberg 等人(2019年) 发现，PDTB 和修辞结构理论-DT 的标注者在 37% 的隐式关系上对关系义项达成一致。Cohen's κ系数主要用于比较单一标签，而这些研究报告的标注者间一致性也基于单一聚合标签。

然而，我们还希望将每个实例获得的 10 条标注与包含多个标签的参考标注进行比较。当存在多个标签时，这种比较就不那么直接，因为一致的概率会被抬高，而且应以不同方式处理部分一致。因此，我们通过完全一致率和部分一致率，以及 Marchal 等人(2022年) 提出的多标签卡帕系数指标，来衡量多个标签之间的标注者间一致性。该指标使用自助法期望一致性对多标签一致性进行校正。对于每个实例，我们考虑众包工作者标注的所有标签，但排除仅获得一票的少数标签[6]。

此外，我们参照现有研究，比较众包标签的分布，使用詹森-香农散度（JSD）（Erk 和 McCarthy，2009年；Nie 等人，2020年；Zhang 等人，2021年）。同样，我们排除仅获得一票的少数标签。由于参考标签中没有分布，因此在与参考标签比较时，我们根据标签的展平分布计算JSD；这意味着，我们将原始投票分布替换为由多于一位标注者投过票的标签所构成的均匀分布。我们将此版本称为JSD flat。

[5] DC 方法中选取的标注人员数量较多，是因为在创建 DiscoGeM 的过程中，DC 标注人员标注了更多数据。

[6] 我们假设每个项目有 10 票，并剔除了得票率低于 20% 的标签，尽管在少数情况下票数可能为 9 票或 11 票。平均而言，被剔除的标签占每个项目票数的 24.8%。

在参考标签中不可用时，与参考标签进行比较，我们根据标签的展平分布计算JSD，也就是说，我们将原始投票分布替换为由多于一位标注者投过票的标签所构成的均匀分布。我们将此版本称为JSD flat。_

作为考察标注者间一致性的第三个视角，我们报告使用 QA/DC 标注的条目上的标注者一致程度。遵循先前的研究（Nie 等人，2020年），我们使用软标签的熵来量化众包标注的不确定性。此处也纳入仅获得一票的标签，因为它们也会影响标注不确定性。计算熵时，我们使用以 $n = 29$为底的对数，其中 n 是可能标签的数量。熵值越低，表明标注者之间的一致性越高，标注标签也越确定。如第1节所述，标注中的分歧可能源于条目、标注者和方法。在同一标注任务中，对特定条目的多次标注若呈现较高熵，则表明该条目存在歧义。

4 结果

我们首先比较两组众包标注之间的标注者间一致性，然后讨论直接分类/问答与参考标注之间的标注者间一致性，最后基于标注不确定性进行分析。此处，实例的“子标签”指所有获得多次标注的关系；“标签分布”指对子标签的投票分布。

4.1 不同方法之间的标注者间一致性

表1 显示，排除仅获得一票的少数标签后，两种方法在每个实例上都产生了两个以上的子标签。这支持了这样一种观点：**多义项标注能够更好地体现这样一个事实：两个话语论元之间往往可能隐含着不止一种语义**。

表 1 还展示了各领域中通过问答和直接分类众包得到的标签之间的标注者间一致性。两种方法之间的一致性较高：两种方法分配的标签（对于多标签标注，至少有一个子标签）在约 88% 的条目中相符。这表明 **两种方法均有效**，因为它们生成了相似的标签集合。

不过，完全一致的得分非常低。这是预料之中的，因为与单标签设置相比，在所有子标签上都匹配的概率也非常低。因此，多标签卡帕系数（将多个标签之间偶然一致的可能性考虑在内）和詹森-香农散度（比较多个标签的分布）更为合适。值得注意的是，我们用于评估的 PDTB 金标准标注并未系统地分配多个关系，双标签的比例也很低。这解释了为什么 PDTB 子集的部分一致率较高，而詹森-香农散度最终表现最差。

4.2 众包标签与参考标签之间的标注者间一致性

表2 比较了通过每种方法众包得到的标签与参考标签；维基百科和 PDTB 子集均提供了参考标签。可以看到，在这两个领域中，两种方法与参考标签之间的完全一致率都高于彼此之间的完全一致率。这表明**两种方法具有互补性**，每种方法都更擅长捕捉不同的语义类型。具体而言，对于维基百科条目，问答方法与参考标签的一致性往往较高；而对于 PDTB 条目，直接分类标注与参考标签的一致性较高。这可能**归因于方法的开发过程**：直接分类方法最初是在 Yung 等人(2019年年使用 PDTB 数据进行测试时开发的，而问答方法则是在 Pyatkin 等人(2020年年使用维基百科和维基新闻数据进行测试时开发的。

4.3 标注不确定性

表 3 比较了两种方法收集的软标签的平均熵。可以观察到，众包工作者所选标签中的不确定性在不同领域间相似，但直接分类的结果始终略低。我们进一步考察了标注不确定性与跨方法一致性之间的相关性，发现方法内熵较低的实例，其方法间的一致性显著更高。同样，我们发现，对于众包过程中熵较低的关系，众包标注与金标准标签的一致性最高。

接下来，我们想检查不同方法和领域中的项目效应是否相似。附录 2 展示了 Wikipedia / PDTB 子集中各项目的标注熵与其参考标注一致度之间的相关性。

	Europarl	小说	维基。	PDTB	all
项目数量	296	304	300	302	1202
QA 子标签/项	2.13	2.21	2.26	2.45	2.21
DC 子标签/项	2.37	2.00	2.09	2.21	2.17
完整/+部分一致	.051/.841	.092/.865	.060/.920	.050/.884	.063/.878
多标签卡帕系数	.813	.842	.903	.868	.857
JSD	.505	.492	.482	.510	.497

表1：直接分类与问答所得标签的比较。完全（或 +部分）一致表示全部（或至少一个子标签）匹配。多标签卡帕系数改编自Marchal 等人（2022年）。詹森-香农散度根据众包子标签的实际分布计算，不包括仅获得一票的标签（数值越小越好）。

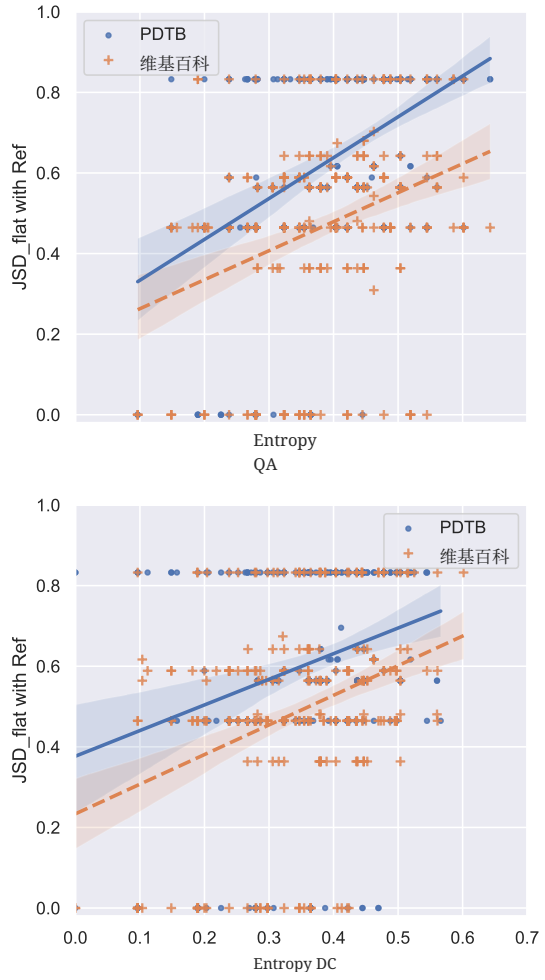


图2：标注熵与众包标签和参考标注之间的詹森-香农散度$_{flat}$ 之间的相关性。

	维基。	PDTB
项目数量	300	302
每项参考子标签数	1.54	1.08
QA：每项子标签数	2.26	2.45
完全/+部分一致	.133/.887	.070/.487
多标签卡帕系数	.857	.449
詹森-香农散度$_{flat}$	.468	.643
DC：子标签/项目	2.09	2.21
完全/+部分一致	.110/.853	.103/.569
多标签卡帕系数	.817	.524
詹森-香农散度$_{flat}$	.483	.606

表2：与问答或直接分类方法的金标准标签进行比较。由于参考子标签的分布不可得，JSD_flat 计算于子标签的均匀分布之间。

	Europarl	维基百科	小说	PDTB
QA	0.40	0.38	0.38	0.41
DC	0.37	0.34	0.35	0.36

表3：按领域划分的 QA/DC 标签分布的平均熵（每种关系有10条标注）。

结果表明，随着标注不确定性增加，两种方法的标注与参考标注之间的差异也随之增大。在 Wikipedia 子集中，不确定性的影响在不同方法之间相似；而在 PDTB 子集中，与 DC 标注相比，问答标注的质量更受不确定性影响。这意味着，方法偏差也存在于标注不确定性层面，因此在例如使用熵作为筛选可靠标注的标准时，应将其纳入考虑。

5 方法偏差的来源

本节中，我们根据各方法收集的义项标签来分析方法偏差。我们

还会考察可能导致这种偏差的方法局限性，并展示如何利用有关方法偏差的信息，通过众包获得更可靠的标签。最后，我们将进行跨领域分析。

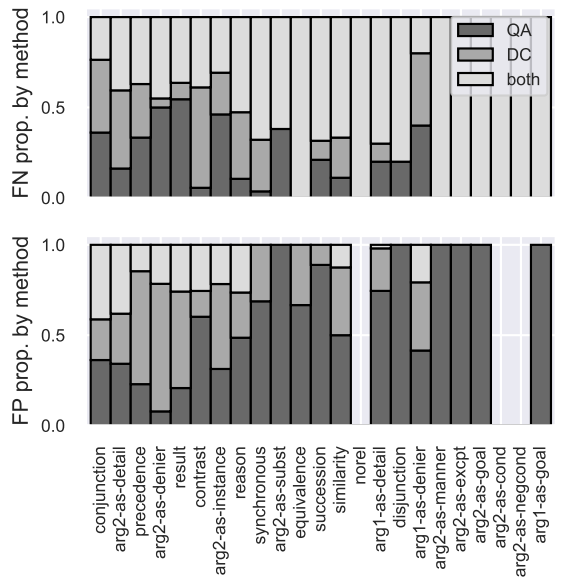


图 3：按方法划分的标注错误分布。将至少由 2 名标注人员标注的标签与维基百科和 PDTB 条目的参考标签进行比较。关系类型按“参考子标签计数”降序排列。

表 5 展示了两种方法针对最常见二级关系所收集标签的混淆矩阵。附录中的图 3 和表4 展示了子标签真阳性和假阳性的分布。这些结果表明，两种方法都偏向某些话语关系。这些偏差的来源可分为两类，我们将在以下小节中详细说明。

5.1 自然语言用于标注的局限性

使用问答方法和直接分类法，以自然语言表示话语关系存在局限性。例如，当问题短语包含连接词时，问答方法会使标注人员感到困惑：[7]

(2) “小家伙，”德思礼先生离开房子时咯咯笑道。**他坐进车里，倒车驶出四号门前**

[7]示例采用以下格式呈现：斜体 = 论元1；粗体 = 论元2；普通文本 = 语境。

标签	FN_{QA}	FN_{DC}	FP_{QA}	FP_{DC}
合取	43	46	203	167
论元2作为细节	42	62	167	152
先行关系	19	18	18	37
论元2作为否定项	38	20	15	47
结果	10	5	110	187
对比	8	17	84	39
论元2作为实例	10	7	44	57
原因	12	17	54	37
同步关系	20	27	11	5
论元2作为替代项	21	13	1	0
等价关系	22	22	2	1
先后关系	17	15	24	3
相似关系	7	8	15	12
无关系	12	12	0	0
论元1作为细节	9	8	39	13
析取	5	4	10	0
论元1作为否定项	3	3	33	31
论元2作为方式	2	2	9	0
论元2作为例外	2	2	1	0
论元2作为目标	1	1	5	0
论元2作为条件	1	1	0	0
论元2作为否定条件	1	1	0	0
论元1作为目标	1	1	3	0

表4：按参考子标签分组的各方法假阴性和假阳性计数

DC / QA	Synch.	Asynch.	Cause	Concession	Contrast	Conjunction	Detail	Instant.	all DC
同步	1	1	1			2	1		6
异步	2	45	17		7	18	7	2	98
因果关系		13	184	5	8	78	36	13	337
让步	2	3	13	14	22	18	22	1	95
对比	1		7	1	21	3	4		37
合取	4	12	37	1	12	184	67	6	323
细节	2	5	40	2	4	52	93	16	214
瞬时			2	1	2	9	10	23	47
全部问答	12	79	301	24	76	364	240	61	

表5：至少由每种关系的2名标注人员标注的最高频二级子标签的混淆矩阵；数值以颜色表示。

开车。 [问答:先后关系、先行关系，直接分类:合取、先行关系]

在上面的例子中，大多数工作**人员都提出了问题**"他离开房子之后呢？"，这很可能是与"**他离开**房子之后做了什么？"的混淆。这可以解释问答中先行关-系和先后关系之间频繁出现的混淆，并导致先后关系的假阳性频繁出现（图3）。[8]

对于直接分类，缺乏常用连接词的罕见关系更难标注，例如：

(3) 他和一只公鸡约好，让它每天早上比其他任何人都早半小时叫醒他；在每天的常规工作开始之前，他还会在任何看起来最需要帮忙的地方出一份志愿劳动。**面对每一个问题、每一次挫折，他的回答都是"我会更加努力！"——他把这句话定为自己的个人格言**。[问答:论元1作为实例；直接分类:结果]

由于缺乏标示这类罕见关系的典型、具体且独立于语境的连接短语，例如"这是.....的一个例子"，因此很难使用直接分类方法标注论元1作为实例关系。相比之下，问答方法允许标注人员反向构造问答对，即以S1作为S2的答案，并使用相同的疑问词，例如：什么是一个例子，可以说明他面对每一个问题时的回答 *[.....]* 都是"我会更加努力！"？。这使标注人员能够标记那些连受训标注者都未能发现的较罕见关系类型。

许多常见的直接分类（DC）存在歧义，例如*but* 和*and*，因此很难消除歧义。为解决这一问题，DC方法会在第二步向标注人员提供没有歧义的连接词。不过，这些无歧义连接词往往相对少见，而且会因其属于并列连词、从属连词还是话语副词而受到不同的句法限制。因此，它们并不适用于所有语境。此外，有些无歧义连接词听起来非常"笨重"，在特定句子中用起来并不自然。例如，*however* 常在第一步中被插入，但它可以标记多种关系，并在第二步中通过从以下选项中选择来消除歧义：*on thecontrary* 用于 对比，*despite* 用于 论元 1- 作为 - 否定项 以及 *despite this* 用于 论元2-作为-否定项。尽管如此 之所以被频繁选用，是因为它适用于大多数语境。这解释了DC方法为何倾向于论元*2*作为否定项 而非 对比（图3：DC产生的大多数假阳性都属于 论元*2*作为否定项，而大多数对比 假阴性也来自DC）。

虽然问答方法也要求标注人员从一组问题开头中进行选择，其中也包含不常见的表达（例如除非什么.....？），但允许标注人员编辑文本，以改进问题的措辞。这有助于减少对更常见问题前缀的偏差影响，也使执行问答任务的众包工作者比执行直接分类任务的工作者更可能选择不常见的关系类型。

5.2 指南规定不充分

Jiang and de Marneffe （2022年）报告称，自然语言推理任务中的一些分歧源于对任务某些方面的定义较为宽泛。我们发现，问答和直接分类也**没有就论元跨度给出足够明确的说明**。话语关系的标注位置是两个连续句子的边界，但这两种方法都没有限制标注人员只能标注恰好跨越这两个句子的话语关系。

更具体地说，问答方法允许众包工作者从其中一个句子中复制部分跨度来构造问题。虽然这样可以确保关系局部存在于两个相邻句子之间，但有时标注人员也会突出显示部分跨度，并标注跨越句子部分内容的关系。例如：

(4) 我赞同皮尔克先生的意见，而且如果我们确实要就卢德福报告投票，这可能是我唯一会赞同他的地方。 **这将是一场有趣的投票**。 [QA:ARG2-AS-DETAIL,REASON; DC:CONJUNCTION,RESULT]

在例(4)中，标注人员构建了这样一个问题：**"关于对 Ludford 报告的投票，哪些内容提供了更多细节？"**。这类似于 PDTB 2.0 和 3.0 的标注手册中的说明，手册规定标注者应选取不必覆盖整个句子的最小跨度。当论元跨度扩展到整个句子时，则应推断其他关系，例如

[8]同样，问题"尽管什么.....？"很容易与"尽管....."混淆，这或许可以解释问答方法中论元*1*作为否定项频繁出现假阳性的原因。

一个结果关系，反映出双方几乎没有共识，这将使投票颇具看点。

通常，一个句子可以被理解为对前一句中某些实体的展开说明。这或许可以解释为什么 ARG1/2-AS-DETAIL 往往会被问答过度标注。图 3显示，与直接分类相比，问答在论元2作为细节上的假阳性计数多出一倍以上；论元1作为细节的差距甚至更大。不过，要过滤掉那些只指向句子某一部分的问题并非易事，因为在某些情况下，被突出显示的实体确实代表了整个论元跨度。[9] 指南中应提供更清晰的说明。

同样，直接分类并不限制标注人员只能标注两个句子之间的关系，请看：

(5)

当同一晶体中存在两个不同掺杂区域时，就会形成半导体结。**这些结处电荷载流子的行为——电荷载流子包括电子、离子和空穴——是二极管、晶体管以及所有现代电子学的基础**。 [Ref:论元2作为细节; QA:论元2作为细节, 合取; DC:合取, 结果]

在这个例子中，许多人插入了*as a re-sult*，这自然标示了句内关系（*...is created as aresult.*）。许多关系可能被错误地标注为结果关系，而这类关系常见于较大的文本片段之间。表格5显示，最常见的混淆是 DC 的因果关系与 QA 的合取。[10] 在二级因果关系义项中，三级结果关系是造成所观察到偏差的主要因素。图3 也显示，结果关系的大多数假阳性都来自 DC 方法。

5.3 基于方法偏差汇总话语关系标注

上述定性分析揭示了标签分布中观察到的一些方法偏差，例如问答方法倾向于将标签赋予 ARG1/2-AS 细节和先后关系，而直接分类倾向于将标签赋予让步和结果。了解这些偏差有助于结合使用不同方法：先用成本效益更高的 DC方法为所有实例标注，再用问答方法重新标注结果关系，因为我们知道直接分类往往会过度标注这类关系。我们在数据中模拟了这一做法，发现维基百科的部分一致程度将从 0.853 提高到 0.913，而 PDTB 的部分一致程度将从 0.569 提高到 0.596。

6 按体裁分析

对于四种体裁（小说、维基百科、Europarl 和《华尔街日报》），我们都有 ~300同时由 DC 和 QA 标注的隐式话语关系。 Scholman 等人(2022年a) 基于 DC方法指出，在 DiscoGeM 中，合取 在维基百科领域中占主导地位，PRECEDENCE 在文学中占主导地位，而 RESULT 在 Europarl 中占主导地位。QA 标注重复了这一发现，如图 4所示。

与其他体裁相比，在 Europarl 中更难取得与多数标签的一致，这也反映在各体裁分布的平均熵中（见表 3）：DC 在 Europarl 领域中的熵最高，而 QA 的熵排名第二（仅次于 PDTB ）。表 1 证实了这些发现，显示两种方法之间的一致性在维基百科中最高，在 Europarl 中最低。

在后一个领域中，直接分类方法产生了更多因果关系：问答法标注为合取的项目中，有36%在直接分类中被标注为结果。[11] 人工检查这些直接分类标注后发现，标注人员只有在 Europarl 子集中才会较频繁地选择考虑到这一点。这一连接短语通常用于标记语用结果关系，其结果解读源自说话者的信念（例(4)）。预计这类关系在言语和论证语境中更为常见，并在宾州话语库3中标记为结果-信念。问答法没有可用于捕捉结果-信念关系类型的问题前缀。因此，与问答法的合取相比，直接分类得到的结果标签更符合宾州话语库3框架。直接分类方法通常会产生更多让步关系，尤其是在 Europarl 中，占 9%，而问答法仅为 3%。另一方面，问答方法似乎更倾向于对比关系，其中大多数（6%）对比关系出现在

[9]例如，本例中的“一些最后的评论”：女士们、先生们，我想最后再谈几点。**这与栖息地指令的实施无关**。[10] 卡方检验证实，观察到的分布与根据偶然分歧所预期的分布显著不同。

[11]这似乎分布在许多标注者之间，因此确实属于方法偏差。

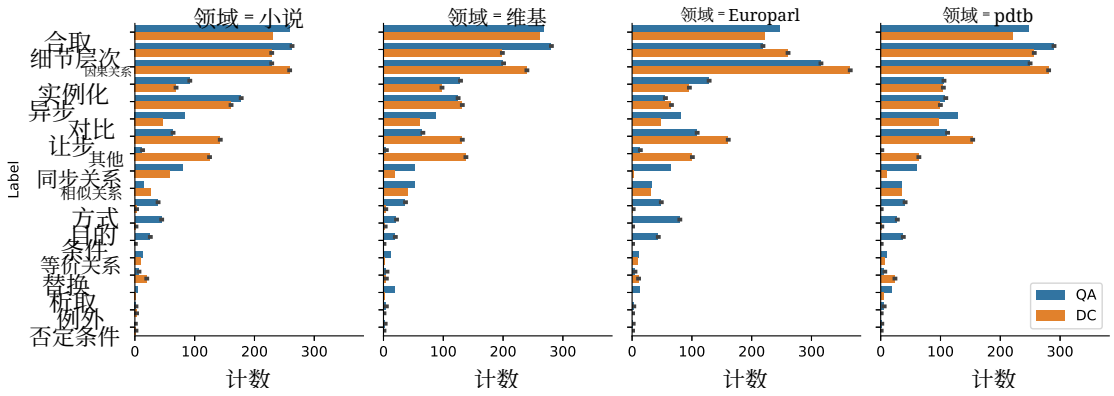


图 4：按领域划分的两种方法所有标注标签的二级子标签计数。

维基百科中，相比之下，直接分类法的比例为 3%。图4还强调，对于问答方法，标注人员倾向于选择更多样的关系类型，而这些类型很少由直接分类法标注，例如目的、条件和方式。

我们得出结论，百科全书类和文学文本最适合采用直接分类或问答进行标注，因为它们的方法间一致性较高（而维基百科文本与金标准的一致性也较高）。另一方面，口语和论证类领域的标注难度更大，因为其中包含更多对关系的语用解读。

7 案例研究：任务设计对话语关系分类模型的影响

对众包标注的分析表明，两种方法具有不同的偏差，并且与领域以及领域中所用语言的风格（也可能包括功能）存在不同的相关性。现在，我们研究任务设计偏差对隐式话语关系自动预测的影响。具体而言，我们开展两项案例研究，以展示任务设计及其造成的标签分布对话语解析模型的影响。

任务与设置 我们将预测隐式话语关系的任务表述如下。模型的输入是两个序列 S_1 和 S_2，它们分别表示某个话语关系的论元。目标是 PDTB 3.0 语义类型（包括第 3 级）。该模型架构与Shi 和 Demberg(2019年提出的隐式话语关系预测模型类似。我们尝试了两种不同的损失和目标：一种是目标为单一多数标签的交叉熵损失，另一种是目标为标注标签概率分布的软交叉熵损失。利用每个实例的 10 条标注，我们为每种关系获得标签分布，并将其用作软目标。研究表明，使用软损失可以提升计算机视觉和自然语言处理任务中的泛化能力（Peterson 等人，2019年；Uma 等人，2020年）。按照Uma 等人(2020年的建议，我们对 30 个可能标签上的语义分布进行归一化[12] ，使用 Softmax。

假设某个关系的标注如下：4 个结果、3 个合取、2 个先后关系、1 个论元1作为细节。对于硬损失，目标是多数标签：结果。对于软损失，我们使用 Softmax 对计数进行归一化（未获得任何标注的标签，其计数均为 0），以获得更平滑且不含零值的分布。

我们使用序列分类设置对 DeBERTa（deberta-base）进行微调（He 等人，2020年），使用 Hugging Face 检查点（Wolf 等人，2020年）。模型训练 30 轮，并采用提前停止策略，批量大小为 8。

数据 除了本研究分析的 1,200 个实例外，我们还将 DiscoGeM 的全部标注用作培训数据。DiscoGeM 使用 DC方法进行标注，新增了 2756 条小说 关系、2504 条欧洲议会 关系和 345 条维基百科 关系。我们

[12]先行关系，论元2作为细节，合取，结果，论元1作为细节，论元2作为否定项，对比，论元1作为否定项，同步关系，原因，论元2作为实例，论元2作为条件，论元2作为替代项，相似性，析取，先后关系，论元1作为目标，论元1作为实例，论元2作为目标，论元2作为方式，论元1作为方式，等价关系，论元2作为例外，论元1作为例外，论元1作为条件，不同连接关系，无关系，论元1作为否定条件，论元2作为否定条件，论元1作为替代项

	PDTB$_{test}$	维基$_{gold}$	TED-M.
DC	0.34†	0.65†	0.36
DC 软	0.29*†	**0.70***†	0.34*
问答+直接分类$_{\cap}$	0.34⋆	0.67	0.37
问答+直接分类$_{\cap}$ 软	0.38*⋆	0.66*	0.31*
问答+直接分类$_{\cup}$	0.35♠	0.49♠	0.36♠
问答+直接分类$_{\cup}$ 软	**0.41***♠	0.67*♠	**0.43***♠

表 6：模型（采用软损失或硬损失）对金标准标签的预测准确率。模型分别使用 DC 数据（DC）、DC 与 QA 的交集（∩）或 DC 与 QA 的并集（∪）进行训练。同一列中的相同符号表示模型间结果存在统计学显著差异（McNemar检验）。

为这些案例研究设计不同的设置。

7.1 案例1：整合来自不同任务设计的数据

本研究旨在考察，基于 DC/QA 方法众包数据训练的模型能否泛化到传统标注的测试集。因此，我们在由专家标注的 300 条维基百科关系（维基百科金标准）、PDTB 3.0 测试集中的所有隐式关系（*PDTB* 测试集），以及 TED-MDB 英语测试集中的隐式关系（Zeyrek 等人，2020年）上进行测试。训练数据有三种选择：1）使用全部 DiscoGeM 标注（**仅 DC**）；2）使用来自所有四个领域的 1200 条 QA 标注，以及 DiscoGeM 其余部分中的 5,605 条 DC 标注（**交集**，∩）；或 3）使用 1200 条合并 QA 和 DC 标签计数（例如，不是 10 个计数，而是 20 个计数）的标注，以及 DiscoGeM 其余部分中的 5,605 条 DC 标注（**并集**，∪）。我们假设，由于标注分布来自更大的样本，这种并集方式会带来更好的结果。在维基百科金标准上进行测试时，训练数据中会移除相应子集的维基百科 关系。我们随机抽取了 30 条关系作为开发集。

结果 表格。 6展示了模型如何泛化到传统标注的数据。在 PDTB 和维基百科测试集上，使用软损失的模型通常优于使用硬损失的模型。另一方面，TED-MDB 中每个关系只有一个标签，因此使用分布式损失进行培训的益处较小。混合直接分类和问答数据，只在软损失情况下提升了 PDTB 的表现。另一方面，合并相应方法的标签计数后，在 PDTB 和 TED-MDB 上都获得了最佳模型性能。在维基百科上，使用仅包含软直接分类分布的数据进行培训时，性能最佳。观察特定标签的性能差异后，我们发现，维基百科测试集上的提升主要来自预测 ARG2-AS-DETAIL 时精确率和召回率的提高；而在 PDTB 上，QA+DC$_{\cap}$ Soft 在预测合取方面表现更好。

我们得出的结论是，使用来自不同任务设计的数据进行训练不会损害性能，而且使用多数投票标签时甚至能略微提升性能。使用分布进行训练时，并集设置（∪）似乎效果最佳。

7.2 案例2：跨领域与跨方法

本研究旨在探究方法偏差如何影响跨领域泛化。换言之，我们希望比较跨领域且跨方法设置与跨领域且同方法设置。我们分别在由 QA 和 DC 标注的 1,200 个实例中的特定领域数据上进行测试，并使用来自 DiscoGeM 的各种领域配置数据（不包括开发集和测试集），以及额外 300 个由 DC 标注的 PDTB 实例进行训练。

表7 展示了本研究采用的不同数据集组合（列），以及域内与跨领域、方法内与跨方法预测的结果（行）。领域变化和标注任务变化都会导致性能下降。有趣的是，结果显示，任务因素对性能的影响比领域更大：使用 DC 分布进行训练时，所有情况下的 QA 测试结果都差于 DC 测试结果。这表明，训练模型时必须考虑任务偏差。总体而言，除小说的域外测试情形外，使用软损失进行训练所获得的泛化准确率与使用硬损失训练相当或显著更高。因此，我们也在 DR 分类中证实了Peterson等人（2019年）和Uma 等人（2020年）的研究发现。

8 讨论与结论

DR标注是一项出了名的困难任务，标注者间一致性较低。标注不仅受编码者的解释影响（Spooren 和 Degand，2010年），也受框架影响（Demberg 等人，2019年）。本研究通过表明任务设计也会对结果产生关键影响，进一步拓展了这些发现——

		除维基 DC 外的所有项			除 EP DC 外的所有项			除 Novel DC 外的所有项			All		
		hard acc	soft acc	soft JSD	hard acc	soft acc	soft JSD	hard acc	soft acc	soft JSD	hard acc	soft acc	soft JSD
Wiki	QA	0.51*	**0.53***	0.38	0.51	0.50	0.37	0.47	0.53*	0.38	0.50*	0.51	0.38
	DC	0.55*	0.59*	0.36	0.60	**0.62**	0.33	0.56	0.61*	0.35	0.58*	0.61	0.35
EP	QA	0.30	0.35	0.43	0.29	0.33	0.42	0.29	**0.38**	0.44	0.29	0.34	0.44
	DC	0.50	**0.53**	0.34	0.41	0.47	0.37	0.49	0.49	0.36	0.47	0.52	0.36
小说	QA	0.44*	**0.47**	0.40	0.41	0.45	0.41	0.40	0.39	0.44	0.39	0.45	0.41
	DC	0.52*	**0.58**	0.33	0.53	0.56	0.34	0.56	0.51	0.39	0.52	0.57	0.36

表7：采用硬损失与软损失进行的跨领域和跨方法实验。列表示训练设置，行表示测试性能。准确率表示预测多数标签的准确率。JDS 比较预测分布（软）与目标分布。* 表示跨方法结果在统计学上不显著（麦内马尔检验）。

果。我们考察了两种不同的众包DR标注任务对所得关系分布的影响。这两种任务的独特之处在于，它们使用自然语言进行标注。尽管这些设计对非专业人士来说更加直观，我们证明，这类基于自然语言的标注设计同样会受到偏差影响，并留有不同解释的空间（传统标注任务也是如此）。

结果表明，两种方法各自具有独特的偏差，但也都有效，因为它们产生的标签集合相似。此外，这些方法似乎具有互补性：两种方法与参考标注的吻合度都高于彼此之间的吻合度。这表明这些方法捕捉到了不同的语义类型。结果还表明，文本领域会使每种方法倾向于产生不同的标签分布。最后，我们模拟了如何根据方法偏差聚合标注，以提高吻合度。

我们建议未来对两种方法进行若干改进。对于问答，我们建议替换以连接词开头的问题前缀选项，例如“After what”。修订后的选项最好以疑问词开头，例如“What happens after..”。这样可以使问题听起来更自然，并有助于避免对三级语义区分产生混淆。对于直接分类，可以改进界面，让标注人员能够高亮论元跨度，从而确认关系存在于两个相邻句子之间。如果允许标注人员对文本进行少量修改，也可以缓解插入某些罕见连接词时遇到的句法限制。

考虑到两种方法都具有优势，也可能存在不足，将它们结合起来用于未来的众包工作可能会很有意思。

鉴于获取 DC 标注的成本更低、速度更快，可以考虑大规模收集 DC 标注，然后对关系义项熵较高的特定子集使用问答方法。另一种选择是将两种方法结合起来：首先让众包工作者插入一个连接词，然后使用问答完成第二步连接词消歧。最后，由于我们证明了通常可能存在不止一个关系义项，因此可以允许标注者为给定关系撰写多个问答对，或插入多个可能的连接词。

DR 分类实验表明，来自不同任务设计的数据之间的泛化很难；在直接分类和问答场景中，这甚至比跨领域泛化更难。此外，我们发现，合并来自不同任务设计的数据分布，有助于提升对第三种来源（传统标注）数据的处理效果。最后，我们证实，采用标签分布的软建模方法可以提高话语分类性能。

任务设计偏差已被认定为标注偏差的来源之一，并且在其他语言学任务中也被视为数据集的一种产物（Pavlick and Kwiatkowski, 2019年；Jiang andde Marneffe，2022年）。我们的研究结果表明，使用通过多种方法收集的数据进行训练，可以减轻此类偏差的影响。其他自然语言处理任务也可能如此，尤其是那些以自然语言形式呈现的任务；比较它们的任务设计或许会成为一个有趣的未来研究方向。因此，我们鼓励研究人员更加关注众包任务设计所引入的偏差。

致谢

本研究得到德国研究联合会资助，资助机构 ID：http://dx.doi.org/10.13039/501100001659，

资助项目：SFB1102：信息密度与语言编码项目，由欧洲研究理事会 ERC-StG 项目（资助编号：677352）和以色列科学基金会（资助编号：2827/21）资助，谨此致谢。我们还感谢 TACL 的审稿人和行动编辑提出的宝贵意见。

参考文献

Rahul Aralikatte、Matthew Lamm、Daniel Hardt 和 Anders Søgaard。2021。将省略解析视为问答：一项评估。载于计算语言学协会欧洲分会第*16*届会议（*EACL*），第810–817页。计算语言学协会。

Lora Aroyo 和 Chris Welty。2013。众包真值：利用众包中的分歧构建关系抽取金标准。*WebSci2013ACM*，2013年（2013）。

Ron Artstein 和 Massimo Poesio。2008年。计算语言学中的编码者间一致性。《计算语言学》，34(4):555–596。

N. Asher。1993年。话语中对抽象对象的指称。Kluwer，诺韦尔，马萨诸塞州，多德雷赫特。

Valerio Basile、Michael Fell、Tommaso Fornaciari、Dirk Hovy、Silviu Paun、Barbara Plank、Mas- simo Poesio 和 Alexandra Uma。2021年。我们需要考虑评估中的分歧。载于*BPPF*。

Samuel Bowman、Gabor Angeli、Christopher Potts 和 Christopher D Manning。2015年。用于学习自然语言推理的大型标注语料库。载于*2015*年自然语言处理实证方法会议论文集，第632–642页。

Samuel Bowman 和 George Dahl。2021年。要修复自然语言理解领域的基准测试，需要做些什么？载于 *2021*年北美计算语言学协会分会会议论文集：人类语言技术，第4843–4855页。

Sven Buechel 和 Udo Hahn。2017a。Emobank：研究标注视角和表示格式对维度情绪分析的影响。载于《欧洲计算语言学协会第*15*届年会论文集：第*2*卷，短篇论文》，第578–585页。Sven Buechel 和 Udo Hahn。2017b。读者、作者与文本：在情感标注中应对文本理解的不同视角。载于《第*11*届语言标注研讨会论文集》，第1–12页。Lynn Carlson 和 Daniel Marcu。2001年。话语标注参考手册。*ISI*技术报告 *ISI-TR-545*，54:1–56。Nancy Chang、Russell Lee-Gold man 和 Michael Tseng。2016年。来自众包的语言学智慧。载于第三届*AAAI*人类计算与众包会议。Tongfei Chen、Zheng Ping Jiang、Adam Poliak、Keisuke Sakaguchi 和 Benjamin Van Durme。2020年。不确定自然语言推理。载于计算语言学协会第*58*届年会论文集，第8772–8779页。John Joon Young Chung、Jean Y Song、Sindhu Kutty、Sungsoo Hong、Juho Kim 和 Walter S Lasecki。2019年。用于估计群体众包答案的高效引出方法。《*ACM*人机交互会议论文集》，3(CSCW):1–25。Jacob Cohen。1960年。名义量表的一致性系数。《教育与心理测量》，20(1):37–46。Alan Cowen、Disa Sauter、Jessica L Tracy 和 Dacher Keltner。2019年。绘制情感图谱：迈向情感体验与表达的高维分类体系。《公共利益中的心理科学》，20(1):69–90。Marie-Catherine De Marneffe、Christopher D Manning 和 Christopher Potts。2012年。事情发生了吗？真实性评估的语用复杂性。《计算语言学》，38(2):301–333。Vera Demberg、Merel CJ Sc holman 和 Fatemeh Torabi Asr。2019年。我们的

话语标注框架兼容性如何？对修辞结构理论标注和PDTB标注进行映射的见解。《对话与话语》，10(1):87–135。

Mark Díaz、Isaac Johnson、Amanda Lazar、Anne Marie Piper 和 Darren Gergle。2018年。解决情感分析中的年龄偏差问题。载于*2018*年计算机系统中的人为因素大会论文集，第 1–14 页。

Anca Dumitrache。2015年。通过众包收集语义标注中的分歧。载于欧洲语义网会议，第 701–710 页。Springer。

Anca Dumitrache、Oana Inel、Lora Aroyo、Benjamin Timmermans 和 Chris Welty。2018年。CrowdTruth 2.0：面向存在分歧的众包质量指标。载于 第一届众包中的主观性、歧义与分歧研讨会，以及第一届探讨众包与偏差管理之间关系的短论文研讨会，*SAD+ Crowd-Bias 2018*，第 11–18 页。CEUR-WS。

Anca Dumitrache、Oana Inel、Benjamin Tim- mermans、Carlos Ortiz、Robert-Jan Sips、Lora Aroyo 和 Chris Welty。2021年。众包真实标注的实证方法。语义网，12(3):403–421。

Yanai Elazar、Victoria Basmov、Yoav Goldberg 和 Reut Tsarfaty。2022年。基于文本的名词短语增强。计算语言学协会会刊，10:764–784。

Katrin Erk 和 Diana McCarthy。2009年。分级词义指派。载于《*2009*年自然语言处理实证方法会议论文集》，第440–449页。

Elisa Ferracane、Greg Durrett、Junyi Jessy Li 和 Katrin Erk。2021年。他们回答了吗？会话话语中的主观行为与意图。载于计算语言学协会北美分会人类语言技术*2021*年会议论文集，第1626–1644页。

Nicholas Fitzgerald、Julian Michael、Luheng He 和 Luke Zettlemoyer。2018年。大规模问答-语义角色标注解析。载于计算语言学协会第*56*届年会论文集

计算语言学（第*1*卷：长篇论文），第2051–2060页。

PengchengHe、Xiaodong Liu、JianfengGao 和 Weizhu Chen。2020年。DeBERTa：采用解耦注意力的增强型 BERT 解码。在国际学习表征会议。

Yufang Hou。2020年。将桥接回指消解视为问答。在计算语言学协会第*58*届年会论文集，第1428–1438页。

Dirk Hovy、Taylor Berg-Kirkpatrick、Ashish Vaswani 和 Eduard Hovy。2013。使用 MACE 学习信任谁。在 *2013*年北美计算语言学协会分会会议：人类语言技术（*NAACL-HLT*）论文集，第1120–1130页，科罗拉多州丹佛。

Christoph Hube、Besnik Fetahu 和 Ujwal Gadi- raju。2019年。理解并缓解众包收集主观判断过程中的工作者偏见。在*2019*年计算机系统中的人为因素大会（*CHI*）论文集，第1–12页。

Terne Sasha Thorn Jakobsen、Maria Barrett、An- ders Søgaard 和 David Lassen。2022年。论证挖掘中标注者偏差对任务定义的敏感性。载于*LREC2022* 第*16*届语言学标注研讨会（*LAW-XVI*）论文集，第44–61页。

Nanjiang Jiang 和 Marie-Catherine de Marneffe。2022年。探究自然语言推理中分歧产生的原因。计算语言学协会会刊，10:1357–1374。

Youxuan Jiang、Jonathan K Kummerfeld 和 Walter Lasecki。2017年。理解众包释义收集中的任务设计权衡。载于计算语言学协会第*55*届年会论文集（第*2*卷：短文），第103–109页。

David Jurgens。2013。拥抱歧义：词义标签标注方法的比较

众包。在*2013*年北美计算语言学协会分会会议：人类语言技术，第556–562页。

Daisuke Kawahara、Yuichiro Machida、Tomohide Shibata、Sadao Kurohashi、Hayato Kobayashi 和 Manabu Sassano。2014年。利用两阶段众包快速构建带有话语关系标注的语料库。载于计算语言学国际会议论文集（*COLING*），第269–278页，爱尔兰都柏林。

YudaiKishimoto、Shinnosuke Sawada、YugoMu- rawaki、Daisuke Kawahara 和 Sadao Kuro- hashi。2018年。改进日语话语关系的基于众包的标注。载于*LREC*。

Wei-Jen Ko、Cutter Dalton、Mark Simmons、Eliza Fisher、Greg Durrett 和 Junyi Jessy Li。2021年。话语理解：一种用于表示句子间联系的问答框架。*arXiv* 预印本 *arXiv:2111.00701*。

Philipp Koehn。2005年。Europarl：用于统计机器翻译的平行语料库。载于*MT Summit X* 论文集，第79–86页，泰国普吉岛。

Yiwei Luo、Dallas Card 和 Dan Jurafsky。2020年。检测媒体对全球变暖的立场。载于计算语言学协会研究成果：*EMNLP 2020*，第3296–3315页。

William C Mann 和 Sandra A Thompson。1988年。修辞结构理论：迈向文本组织的功能理论。《文本：话语研究跨学科期刊》，8(3)：243–281。

Christopher D Manning。2006年。局部文本推理：它难以界定，但见到时你便知道——自然语言处理需要它。

Marian Marchal、Merel Scholman、FrancesYung 和 Vera Demberg。2022年。确立多标签标注的标注质量。载于《第*29*届国际计算语言学会议论文集》，第3659–3668页。

Sewon Min、Julian Michael、Hannaneh Hajishirzi 和 Luke Zettlemoyer。2020年。AmbigQA：回答有歧义的开放领域问题。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第5783–5797页。

Yixin Nie、Xiang Zhou 和 Mohit Bansal。2020年。我们能从人们对自然语言推理数据的集体意见中学到什么？载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第9131–9143页。

Rebecca J Passonneau 和 Bob Carpenter。2014年。标注模型的益处。《计算语言学协会会刊》，2:311–326。

Ellie Pavlick 和 Tom Kwiatkowski。2019年。人类文本推理中固有的分歧。《计算语言学协会会刊》，7:677–694。

Joshua C Peterson、Ruairidh M Battleday、Thomas L Griffiths 和 Olga Russakovsky。2019年。人类的不确定性使分类更加稳健。载于*IEEE/CVF*计算机视觉国际会议论文集，第9617–9626页。

Barbara Plank、Dirk Hovy 和 Anders Søgaard。2014年。是语言学上存在争议，还是单纯错误？载于第*52*届计算语言学协会年会论文集（第*2*卷：短文），第507–511页。

Massimo Poesio 和 Ron Artstein。2005年。重新审视照应标注的可靠性：将歧义纳入考量。载于语料库标注前沿*II*研讨会论文集：空中楼阁，第76–83页。

Massimo Poesio、Patrick Sturt、Ron Artstein 和 Ruth Filik。2006年。欠明确化与照应：理论问题及初步证据。话语过程，42(2):157–175。

Vinodkumar Prabhakaran、Aida Mostafazadeh Davani 和 Mark Diaz。2021年。关于发布标注者级别的标签和数据集信息。载于第*15*届联合语言标注研讨会（*LAW*）和第*3*届

意义表示设计（*DMR*）研讨会，第133–138页。

Valentina Pyatkin、Ayal Klein、Reut Tsarfaty 和 Ido Dagan。2020年。《QADiscourse——将话语关系表示为问答对：表示、众包与基线》。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第2804–2819页。

Pranav Rajpurkar、Robin Jia 和 Percy Liang。2018年。《了解自己不知道什么：SQuAD中的不可回答问题》。载于第*56*届计算语言学协会年会论文集（第*2*卷：短文），第784–789页。

Ines Rehbein、Merel Scholman 和 Vera Demberg。2016年。口语中话语关系的标注：PDTB与CCR框架的比较。载于第十届国际语言资源与评测会议（*LREC'16*）论文集，第1039–1046页，斯洛文尼亚波多罗日。欧洲语言资源协会（ELRA）。

Stefan Riezler。2014年。论实证计算语言学中的理论术语问题。《计算语言学》，40(1):235–245。

Hannah Rohde、Anna Dickinson、Nathan Schneider、Christopher Clark、Annie Louis 和 Bonnie Webber。2016年。填补理解话语副词时的空白：众包诱导任务中的一致性、冲突与语境依赖性。载于第十届语言学标注研讨会论文集（*LAW X*），第49–58页，德国柏林。

Ted J. M. Sanders、WilbertP. M. S. Spooren 和 Leo G. M. Noordman。1992年。迈向连贯关系的分类体系。《话语过程》，15(1):1–35。

Merel C J Scholman 和 Vera Demberg。2017年。众包话语解释：语境的影响及连接词插入任务的可靠性。载于第十一届语言学标注研讨会论文集（*LAW*），第24–33页，西班牙瓦伦西亚。

Merel C. J. Scholman、Tianai Dong、Frances Yung 和 Vera Demberg。2022年。Discogem：一个混合体裁隐式话语关系众包语料库。载于第十三届语言资源与评估国际会议（*LREC'22*）论文集，法国马赛。欧洲语言资源协会（ELRA）。

Merel C. J. Scholman、Valentina Pyatkin、Frances Yung、Ido Dagan、Reut Tsarfaty 和 Vera Dem- berg。2022b。《众包话语关系标注中的设计选择：工作者筛选和培训的影响》。载于《第十三届语言资源与评估国际会议（*LREC'22*）论文集》，法国马赛。欧洲语言资源协会（ELRA）。

Wei Shi 和 Vera Demberg。2019年。使用序列到序列网络学习显化连接词，以进行隐式话语关系分类。载于第十三届计算语义学国际会议论文集——长篇论文，第188–199页，瑞典哥德堡。计算语言学协会。

Rion Snow、Brendan O'Connor、Daniel Jurafsky 和 Andrew Y Ng。2008年。廉价又快速——但效果好吗？评估自然语言任务中的非专家标注。载于自然语言处理实证方法会议（*EMNLP*）论文集，第254–263页，威基基，夏威夷州。

Wilbert P. M. S. Spooren 和 Liesbeth Degand。2010年。连贯关系编码：可靠性与效度。《语料库语言学与语言学理论》，6(2):241–266。

Alexandra Uma、Tommaso Fornaciari、Dirk Hovy、Silviu Paun、Barbara Plank 和 Massimo Poe- sio。2020年。软损失函数的理由。载于*AAAI*人类计算与众包会议论文集，第8卷，第173–177页。

Alexandra N Uma、Tommaso Fornaciari、Dirk Hovy、Silviu Paun、Barbara Plank 和 Massimo Poesio。2021年。从分歧中学习：综述。人工智能研究杂志，72:1385–1470。

Zeerak Waseem。2016年。你是种族主义者，还是我看错了？标注者对 Twitter 仇恨言论检测的影响。载于论文集

首届自然语言处理与计算社会科学研讨会，第138–142页。

Bonnie Webber。2009年。宾夕法尼亚树库中话语的体裁区分。在*ACL*第*47*届年会与第*4*届*AFNLP*自然语言处理国际联合会议论文集，第674–682页。

Bonnie Webber、Rashmi Prasad、Alan Lee和Aravind Joshi。2019年。《宾夕法尼亚话语树库3.0标注手册》。费城，宾夕法尼亚大学。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz等。2020年。Transformers：最先进的自然语言处理。在*2020*年自然语言处理实证方法会议论文集：系统演示，第38–45页。

Frances Yung、Vera Demberg 和 Merel Schol- man。2019年。通过两步连接词插入任务众包话语关系标注。载于第*13*届语言标注研讨会论文集，第16–25页。

Deniz Zeyrek、Amália Mendes、Yulia Grishina、Murathan Kurfalı、Samuel Gibbon 和 Maciej Ogrodniczuk。2019年。TED多语言篇章库（TED-MDB）：采用PDTB风格标注的平行语料库。语言资源与评估，第1–27页。

Deniz Zeyrek、Amália Mendes、Yulia Grishina、Murathan Kurfalı、Samuel Gibbon 和 Maciej Ogrodniczuk。2020年。TED多语言篇章库（TED-MDB）：采用PDTB风格标注的平行语料库。语言资源与评估，54卷2期：587–613。

Shujian Zhang、ChengyueGong 和 Eunsol Choi。2021年。使用不同数量的标注进行学习：从零个到多个标签。载于*2021*年自然语言处理实证方法会议论文集，第7620–7632 页。

Šárka Zikánová、Jiˇrí Mírovsk`y 和 Pavlína Synková。2019年。《布拉格篇章树库中的显性与隐性篇章关系》。载于《文本、言语与对话：第*22*届国际会议，*TSD 2019*，卢布尔雅那，斯洛文尼亚，*2019*年*9*月*11*日至*13*日，论文集》第*22*卷，第*236–248*页。Springer。