arXiv:2302.00762v2 [cs.CL] 3 Feb 2023

AMBICOREF：评估人类和模型对歧义共指的敏感性

YueweiYuan 、**ChaitanyaMalaviya** 和
MarkYatskar宾夕法尼亚大学 {yuewei, cmalaviy,
myatskar}@seas.upenn.ed u

摘要

给定句子“Abby 告诉 Brittney，她让 Courtney 难过了”，人们会难以理解“她”指的是谁，并会要求澄清。然而，如果将“让……难过”替换为“拥抱”，那么“她”毫无歧义地指向 Abby。我们研究现代共指消解模型是否会对这种代词歧义保持敏感。为此，我们构建了**AMBICOREF**，这是一个诊断语料库，包含由歧义和无歧义指称对象构成的最小句对。我们的示例将心理语言学研究推广到人类对特定动词及其论元排列所引发歧义的感知。分析表明：(1) 在 AmbiCoref 示例中，人类对歧义指称对象的判断确信程度低于对无歧义指称对象的判断；(2) 大多数共指模型在歧义句对和无歧义句对之间的输出差异很小。我们发布 AMBICOREF 作为诊断语料库，用于测试模型对歧义的处理方式是否与人类相似。[1]

1 引言

歧义是语言Wasow 等人，2003中的一个基本特征，一些语言学家认为，这种特征源于高效沟通的压力（Haywood 等人，2005；Piantadosi 等人，2012）。近期，多项研究强调了问答（Min 等人，2020；Guo 等人，2021）、框架消歧（Dumitrache 等人，2019）、照应消解（Poesio 和 Artstein，2005）和语言建模（Aina 和 Linzen，2021）等任务中歧义的存在。然而，目前尚缺乏对模型在多种语言处理问题中如何应对歧义进行系统评估的研究。我们针对共指消解开展了这样一项研究。

共指消解对于自然语言理解至关重要，尤其是在长上下文中，例如对话。对话中可能自然地产生歧义，但现有模型对于此类共指并没有明确定义的目标行为。相比之下，人们遇到共指歧义时能够识别出来，并可以请求澄清。现有资源（如 OntoNotes（Weischedel 等人，2013））并未对这类实例提供细粒度标注，无法用于评估模型行为。这可能导致模型未能校准到适当处理对含歧义陈述的不同解释所带来的不确定性。在这项工作中，我们探究：基于这些资源训练的模型对歧义有多敏感?

为了理解现有共指消解模型如何应对歧义，我们构建了一个诊断语料库，即**AMBICOREF**。AMBICOREF 由包含歧义和无歧义指称对象的最小对组成，这些最小对由四类模板生成。我们通过将上下文长度缩减为一个句子，并构造参与其中的动词对其论元的解释约束不足的句子来实现歧义。例如，在表1第 2 行中，我们的第一个模板利用了表达主观体验的动词所产生的歧义。[2] 这些模板参考心理语言学研究（Springston，1976；Caramazza 等人，1977；Rohde 和 Kehler，2014）设计而成，而我们的核心贡献之一，是将这些研究中的观察推广开来，以创建数千个实例。为此，我们识别出 VerbNet 中可能包含合适动词的类别，并将其手动分配给模板。再结合通过名词列表引入的多样性，AMBICOREF 包含超过 9.6 万个句子。

我们通过众包收集判断，验证人类是否会以预期的方式感知 AMBICOREF 中的实例（§3）。我们请标注者找出句子中代词的共指对象，并评定自己的置信度，以体现

[1]我们的数据集和代码可在 https://github.com/LucyYYW/AmbiCoref.

[2]这类实例需要特定的句法安排：如果将代词移至 bored 的宾语位置，第 2 行中的歧义实例就不再有歧义。

	Type	歧义	模板	数量
1	体验者宾语 (ECO-1)	✗	$[Emily]_A$ 告诉 $[Jessica]_B$ 她说 [她]$_A$ [看见了] [Brian]。	11336
2	体验者宾语 (ECO-1)	✓	$[Emily]_A$ 告诉 $[Jessica]_B$， [她]$_?$ [让] [Brian感到厌烦]。	11336
3	体验者宾语（ECO-2）	✗	$[The\ mother]_A$ 告诉 $[the\ sister]_B$ 说 [她]$_A$ [看见了] 这位客户。	11336
4	体验者宾语（ECO-2）	✓	$[The\ mother]_A$ 告诉 $[the\ sister]_B$ 说 [她]$_?$ [让] 这位客户感到厌烦。	11336
5	体验者主语（ECS-1）	✗	$[The\ aunt]_A$ 告诉 $[Sarah]_B$ 说 [女儿] [与.....见面] [她]$_A$。	4472
6	体验者主语 (ECS-1)	✓	$[The\ aunt]_A$ 告诉 $[Sarah]_B$ 说 [女儿] [喜欢] [她]$_?$。	4472
7	体验者主语 (ECS-2)	✗	$[The\ father]_A$ 告诉 $[the\ son]_B$ 客户说 [与.....见面] [他]$_A$。	4472
8	体验者主语 (ECS-2)	✓	$[The\ father]_A$ 告诉 $[the\ son]_B$ 客户 [喜欢] [他]$_?$。	4472
9	隐含因果性（IC）	✗	$[Abby]_A$ [打电话] $[Jane]_B$ 因为 [她]$_A$ [想道歉]。	8424
10	隐含因果性（IC）	✓	$[Abby]_A$ [打电话] $[Jane]_B$ 因为 [她]$_?$ [很快就要离开了]。	8424
11	转移（TOP）	✗	$[Daniel]_A$ [烤制] $[the\ boy]_B$ [蛋糕] [之后] [他]$_B$ [要了一块]。	8424
12	转移（TOP）	✓	$[Daniel]_A$ [烤制] $[the\ boy]_B$ [蛋糕] [之前] [他]$_?$ [吃了午餐]。	8424

表1：构成 AMBICOREF 的六组模板对概述。模板槽位用方括号标出，簇以角标和颜色标记。所有模板都将一个无歧义句子与一个有歧义句子配对，两者的区别仅在于动词短语的选择。

歧义判断的渐变性（Schutze1995）。我们发现，对于无歧义实例，人们会高度认定代词指向预期的名词；而对于有歧义的实例，人们在所有模板中的置信度都较低，大多数参与者要么缺乏信心，要么将其标记为有歧义。这表明，人们以本质上不同的方式处理
AMBI-COREF 中有歧义和无歧义的句子。

AMBICOREF 可用于评估模型在存在歧义时的行为。我们分析了五种具有代表性的英语模型：CoreNLP 中的三种模型（Manning 等人，2014）、SpanBERT（Joshi 等人，2020）和 NeuralCoref 4.0（Wolf 等人，2020）（§4）。我们的主要评估比较歧义和无歧义样本中代词的共指簇分配。我们分析的 5 种模型中有 4 种几乎没有表现出行为变化。与人类不同，共指模型在存在歧义时基本不会改变其判断。我们的分析表明，要在面对有歧义的输入时实现类似人类的行为，模型可能需要明确考虑歧义。

2 数据集构建

为了理解模型对共指歧义的敏感度，我们构建了 AMBICOREF，使用四种类型的模板，如表 1 所示。这些模板以最小对的形式创建，歧义与无歧义版本之间的唯一区别在于动词短语的选择。请注意，尽管歧义是一个程度性现象，我们将“歧义”一词用于那些 更有可能引发人类判断者产生歧义的情况，反之亦然。动词短语从 VerbNet 中适用的动词类别提取（Schuler，2005），这些类别通过对 VerbNet 聚类进行人工标注来确定。[3] 每个模板都通过填入动词、姓名、名词短语和符合性别的代词进行实例化，大幅扩展了先前研究中所识别案例的多样性。

2.1 模板类型

宾语体验者约束（ECO） Springston （1976）提出了适用于补语结构的体验者约束，我们在模板中将其操作化。将宾语标记为某种情绪的体验者的动词，会限制将宾语位置代词指派给陈述性交流动词主语的可能性。相反，当代词位于体验者动词的主语位置时，这种指派不受限制。例如，在表1的第2行中，“bored”的主语位置中的代词存在歧义（但若位于宾语位置则不会产生歧义）。如果主要动词不施加体验者约束，即第1行的情况，那么主语位置中的代词就是无歧义的。我们设置了两个变体：使用姓名的变体（第1、2行）和使用一般实体的变体（第3、4行）。

主语体验者约束 (ECS)体验者约束还表明，将主语标记为情绪体验者的动词会限制主语位置代词的指派。代词位于

[3]我们考察了以下动词类别中的动词：31：心理动词（心理状态动词）；13：所有权变化动词；37：交流动词，因为它们在概念上与歧义所需的条件高度吻合。作者分别评估了各簇中的动词，以确定其是否适合用于模板。

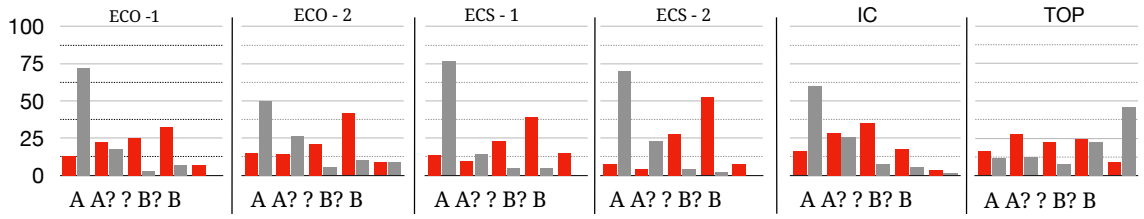

图1：对有歧义（ ）和无歧义（ ）句子的人工标注。我们根据标注者识别的是名词 A 还是名词 B，以及其标注为确定还是可能（以 ? 标记），对人工标注进行简写。例如，A? 表示名词 A，可能。无歧义实例的真实标注从左到右依次为 A、A、A、A、A、B。标注者按预期理解无歧义示例，并降低对有歧义示例的置信度。

宾语位置时，其指派不受限制。例如，在表 1第 6 行中，当代词放在宾语位置时（而非主语位置时），“喜欢”存在歧义。我们构造了使用姓名（第 5、6 行）和实体（第 7、8 行）的变体。

隐含因果性 (IC) Caramazza 等人(1977) 提出，动词的隐含因果性可以决定代词指派的方向。例如，在表1 第 9 行中，短语“想要道歉”建立了“Emily 打了电话”的原因，因此代词指向“打电话”的主语。相反，在第 10 行中，短语“很快就要离开”未能建立这种关系，因此代词存在歧义。对于这些模板（第 9、10 行），我们手动更改涉及实体的名称，并将动词（即 called ）与暗示因果关系的自拟短语（即 apologizing）配对。

所有权转移（TOP）Rohde 和 Kehler （2014）指出，在“John 把漫画递给了 Bill。 他 ……”这类所有权转移语境中，代词指向主语和非主语的可能性相同。我们依据这一观察，围绕涉及来源—目标所有权转移的动词创建模板。我们将示例精简为一个句子，并将转移事件与一个原因配对。例如，在表格 1 第 11 行中，“asked for one”（要了一个）这一短语将代词限定为指向“bake”（烘焙）的接收者。相反，“午餐前”并不会施加这种限制，因为事件发生前，接收者或给予者都可能“吃过午餐”。模板会改变姓名、动词、物品、原因和介词（第 11、12 行）。

2.2 填充模板槽位

对于每个模板，我们都会构建一个合适的动词短语列表、理由列表（适用于 IC 和 TOP 模板），以及一个共享的性别化姓名和名词短语。动词短语是通过人工检查 VerbNet 类别构建的。为控制姓名偏见，我们从常用姓名列表中随机抽取姓名[4] 这些姓名来自过去 50 年，并复用了 Wino-Bias 中的性别化名词短语列表（Zhao 等人，2018）。排除姓名和名词短语的变体后，ECO、ECS、IC 和 TOP 的模板分别有 114、45、81 和 82 个实例。

3 人工判断

用于创建 AMBICOREF 通过词汇资源推广了多项心理语言学研究。接下来，我们验证人们是否会以预期的方式感知这些示例中的歧义。我们针对每个模板提取一部分数据，并询问 Amazon Mechanical Turk 工作人员代词指向哪个人（在表格 *1* 中标为 *A* 或 B），并评定置信度（确定或 可能）。标注者还可以将指称对象标记为完全歧义。每个模板和动词槽位均匀随机抽取一个句子。每个实例收集了 3 条标注。[5] 有关人工判断收集的详情，请参阅附录 A 。

图 1总结了我们的结果。对于无歧义模板，人类判断倾向于预期的共指判定。对于无歧义的 ECO、ECS、IC 和 TOP 实例，预期解读分别有 83.2%、91.9% 和 85.8%、68.3% 的概率被选择为可能或确定。对于有歧义的实例，标注明显转向歧义。如先前的研究所示，人类在有歧义的实例上表现出显著的分歧（Poesio 等人，2019）。这

[4]https://www.ssa.gov/oact/babynames/decades/[5]在存在歧义的情况下，标注者无法可靠地标注某个特定类别，往往会以较低的置信度猜测。因此，我们不仅报告每个实例的多数意见，而是直接报告每个句子的多条标注，以观察总体趋势。

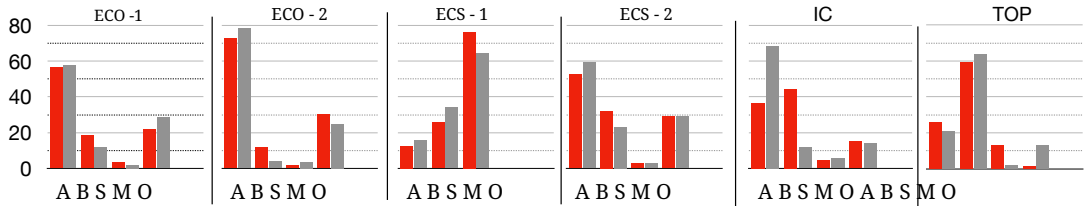

图 2：对于基于 SpanBERT 的模型，在所有模板中，歧义实例（ ）和无歧义实例（ ）分别落入我们五种情形的比例。其他模型的变化都微乎其微（红色和灰色分布几乎相同）。从左到右，无歧义实例的真实标注对应 A、A、A、A、A、B。

一点体现在许多模板中，例如 TOP，在这些模板中，人类给出的回答几乎均匀分布。

4 模型评估

现在，我们通过在 A 上进行评估，考察能否检测出现有共指消解模型对歧义的敏感性MBICOREF。我们使用五种有代表性的模型进行实验 [6] ：Hugging Face 的 NeuralCoref 4.0 模型[7]、独立框架中用于端到端共指消解的 SpanBERT 表征（Joshi 等人，2020），以及 Stanford CoreNLP 中的三种模型（Joshi 等人，2019）：确定性模型（Manning 等人，2014）、统计模型（Lee 等人，2013）以及神经提及排序模型（Clark 和 Manning，2015）。所有模型均在 CoNLL 2012 数据集（Clark 和 Manning，2016）上训练。

在这里，我们评估模型的最终预测，而非其在各种可能选项上的分布。原因有二：（1）并非所有模型都会生成分布；（2）初步分析发现，和其他设置中一样，模型存在校准失配（De- sai 和 Durrett，2020；Jiang 等人，2021），因此直接解读其输出分数并不可靠。

4.1 设置

本节探讨：模型处理相似的无歧义示例和有歧义的示例时，方式是否存在差异？由于我们的示例是合成生成的，因此将无歧义示例用作一种对照。如果模型在至少 40% 的无歧义示例中无法将代词与正确的名词关联起来，我们便会在评估时省略该模板。

我们通过将模型行为分解为若干情形来进行分析，这些情形涵盖了单个句子中代词的所有可能簇标注任务。我们计算模型输出包含以下内容的簇的比例：

[6]约需一周的连续 Colab GPU 计算。[7] https://github.com/huggingface/neuralcoref

模型	平均 EMD %	模板
SpanBERT	11.7	5
CoreNLP 神经网络型	3.5	5
NeuralCoref 4.0	4.0	5
CoreNLP 统计型	1.2	3
CoreNLP 确定型	0.6	5

表 2：匹配的歧义和无歧义情形分布之间的平均推土机距离，以及模型至少能正确判断 40% 无歧义情形的模板数量。

- 情形**A**：代词和名词 A
- 情形**B**：代词和名词 B
- 情形**S**：代词作为单例
- 情形**M**：代词、名词 A 和名词 B
- 情形**O**：代词和任何其他文本跨度

例如，图 2 展示了 SpanBERT 对每个模板在这些情形下的输出分布。对于模型性能高于阈值的每个此类分布，我们使用推土机距离（EMD）比较歧义（红色条）和无歧义（灰色条）分布（Pele 和 Werman，2009 年）[8]。表 2报告了高于阈值的模板数量及其平均 EMD。

4.2 结果

总体而言，我们评估的大多数模型在歧义模板与无歧义模板下的输出分布几乎没有变化，EMD 接近于零便证明了这一点。大多数模型在六个模板中的五个上接受评估，但 TOP 往往被排除在外，因为对大多数系统而言，TOP 本身就是一个难以处理的无歧义案例。

在我们评估的模型中，只有 SpanBERT 在面对歧

[8]推土机距离表示使两个概率分布相匹配所需的概率质量。因此，与 KL 散度等其他可能的度量相比，它有助于我们以更直观的方式比较歧义和无歧义实例的分布。

义输入时表现出显著的行为偏差。图 2详细展示了 SpanBERT 在各个模板上的表现。尽管其平均 EMD 高于其他模型，但预测结果总体上仍未发生变化。当决策发生变化时，代词通常会与另一个名词关联。例如，在 ECO-1 的歧义案例中，SpanBERT 减少了合并输出，转而更频繁地将代词与名词 B 关联。在歧义案例中，其他模型总体上会将第一个名词短语（A）与代词关联。

5 讨论与结论

总体而言，我们的结果表明，模型的行为与人类处理歧义共指的方式存在显著偏差。这进一步证明，模型未能把握人们理解语言的某些方面，尤其是在话语层面（Upadhye 等人，2020）。这可能部分是因为模型是在未考虑判断分布的资源上训练的。因此，当出现歧义时，模型的行为缺乏明确定义，且校准效果不佳。

使用更细粒度的共指判断来训练模型，或许能使模型更好地贴合人类行为。改进模型校准的技术也可能有效，使模型在出现歧义时能够 abstain 或寻求澄清。我们希望 AMBICOREF 能作为一种诊断集，供未来的建模方法用于评估其对语言中歧义实例的敏感度。

6 局限性

我们的研究完全聚焦于英语中的共指，所用模型均在高资源条件下训练。此外，我们识别出的歧义案例具有英语特性，而我们插入模板中的名字都是常见的美国人名。我们的研究结果能否推广到低资源的非美国英语环境，仍是一个尚待解答的问题。

我们用来评估模型的语言具有模板化特征。尽管我们努力考虑到不自然的数据，但如果只评估模型擅长处理的模板，模型仍难以彻底解决所有无歧义示例。这给未来的模型构建者带来了挑战。另一方面，我们的模板可能无法反映模型将接受测试的某种特定现实世界分布。

致谢

我们感谢 Chris Callison-Burch 和 PennNLP 团队为这项工作提出的宝贵意见。

参考文献

Laura Aina 和 Tal Linzen。2021。语言模型理解了提示存在歧义：通过生成探测句法不确定性。载于第四届分析与解读自然语言处理神经网络 *BlackboxNLP* 研讨会论文集，第 42–57 页，蓬塔卡纳，多米尼加共和国。计算语言学协会。

Alfonso Caramazza、Ellen Grober、Catherine Garvey 和 Jack Yates。1977。照应代词理解。《言语学习与言语行为杂志》，16(5)：601–609。

Kevin Clark 和 Christopher D Manning。2015。基于实体的模型堆叠共指消解。载于第 *53* 届计算语言学协会年会暨第 *7* 届国际自然语言处理联合会议论文集（第 *1* 卷：长篇论文），第 1405–1415 页。

Kevin Clark 和 Christopher D. Manning。2016。用于提及排序共指模型的深度强化学习。载于*2016* 年自然语言处理实证方法会议论文集，第 2256–2262 页，得克萨斯州奥斯汀。计算语言学协会。

Shrey Desai 和 Greg Durrett。2020。预训练 Transformers 的校准。载于*2020* 年自然语言处理实证方法会议（*EMNLP*）论文集，第 295–302 页，线上。计算语言学协会。

Anca Dumitrache、Lora Aroyo 和 Chris Welty。2019。带有歧义的众包框架消歧语料库。载于计算语言学协会北美分会 *2019* 年会议论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），第 2164–2170 页，明尼苏达州明尼阿波利斯。计算语言学协会。

Meiqi Guo、Mingda Zhang、Siva Reddy 和 Malihe Alikhani。2021 年。Abg-coqa：澄清对话式问答中的歧义。载于第三届自动化知识库构建会议。

Sarah L Haywood、Martin J Pickering 和 Holly P Branigan。2005 年。说话者在对话过程中会避免歧义吗？《心理科学》，16(5)：362–366。

ZhengbaoJiang、JunAraki、HaiboDing 和 Graham Neubig。2021 年。我们如何知道语言模型何时真正掌握了知识？论问答语言模型的校准。《

计算语言学协会会刊，9：962–977。

Mandar Joshi、Danqi Chen、Yinhan Liu、Daniel S. Weld、Luke Zettlemoyer 和 Omer Levy。2020。SpanBERT：通过表示和预测文本跨度改进预训练。计算语言学协会会刊，8：64–77。

Mandar Joshi、Omer Levy、Luke Zettlemoyer 和 Daniel Weld。2019。用于共指消解的 BERT：基线与分析。载于*2019* 年自然语言处理实证方法会议暨第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第 5803–5808 页，中国香港。计算语言学协会。

Heeyoung Lee、Angel Chang、Yves Peirsman、Nathanael Chambers、Mihai Surdeanu 和 Dan Jurafsky。2013。基于以实体为中心、按精确率排序的规则进行确定性共指消解。计算语言学，39(4)：885–916。

Christopher Manning、Mihai Surdeanu、John Bauer、Jenny Finkel、Steven Bethard 和 DavidMcClosky。2014 年。Stanford CoreNLP 自然语言处理工具包。载于第 *52* 届计算语言学协会年会论文集：系统演示，第 55–60 页，马里兰州巴尔的摩。计算语言学协会。

Sewon Min、Julian Michael、Hannaneh Hajishirzi 和 LukeZettlemoyer。2020 年。AmbigQA：回答开放领域中的歧义问题。载于*2020* 年自然语言处理实证方法会议论文集（*EMNLP*），第 5783–5797 页，线上。计算语言学协会。

Ofir Pele 和 Michael Werman。2009 年。快速且稳健的推土机距离。载于*2009* 年 *IEEE* 第 *12* 届国际计算机视觉会议，第 460–467 页。IEEE。

Steven T. Piantadosi、Harry Tily 和 Edward Gibson。2012。语言中歧义的交际功能。《认知》，122(3)：280–291。

Massimo Poesio 和 Ron Artstein。2005。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*：空中楼阁》研讨会论文集，第 76–83 页，美国密歇根州安娜堡。计算语言学协会。

Massimo Poesio、Jon Chamberlain、SilviuPaun、Juntao Yu、Alexandra Uma 和 Udo Kruschwitz。2019。关于照应解释的多重判断与分歧的众包语料库。载于*2019* 年北美计算语言学协会年会论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），第 1778–1789 页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Sameer Pradhan、Alessandro Moschitti、Nianwen Xue、Olga Uryupina 和 Yuchen Zhang。2012 年。CoNLL-2012 共享任务：在 OntoNotes 中对多语言无限制共指进行建模。载于 *EMNLP* 与 *CoNLL* 联合会议——共享任务，第 1–40 页，济州岛，韩国。计算语言学协会。H. Rohde 和 A. Kehler。2014 年。语法和信息结构对代词产出的影响。《语言、认知与神经科学》，29(8):912–927。Karin Kipper Schuler。2005 年。*VerbNet*：广覆盖、全面的动词词库。宾夕法尼亚大学。Hinrich Schutze。1995 年。语言学习中的歧义：计算与认知模型。斯坦福大学。F. Springston。1976 年。照应代词理解中的动词衍生约束。发表于东部心理学会（1976 年）。Shiva Upadhye、Leon Bergen 和 Andrew Kehler。2020 年。预测指称：语言模型学会了哪些关于话语模型的内容？载于*2020* 年自然语言处理实证方法会议（*EMNLP*）论文集，第 977–982 页，线上。计算语言学协会。Thomas Wasow、Amy Perfors 和 David Beaver。2003 年。歧义之谜。Ralph Weischedel、Martha Palmer、Mitchell Marcus、Eduard Hovy、Sameer Pradhan、Lance Ramshaw、Nianwen Xue、Ann Taylor、Jeff Kaufman、Michelle Franchini 等。2013 年。OntoNotes 5.0 版 ldc2013t19。语言数据联盟，宾夕法尼亚州费城，23。Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Remi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020 年。Transformers：最先进的自然语言处理。载于*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，线上。计算语言学协会。Jieyu Zhao、Tianlu Wang、Mark Yatskar、Vicente Ordonez 和 Kai-Wei Chang。2018 年。共指消解中的性别偏见：评估与去偏方法。

人工判断测试

在我们所有的人工判断测试中，我们要求标注者主要来自英语国家：美国、英国、加拿大或澳大利亚。此外，标注者需要至少完成 1000 个获批准的人类智能任务（HIT），且 HIT 接受率至少为 98%。每个 HIT 包含 10 个示例，我们估计每个 HIT 的完成时间为~5 分钟，因此每个 HIT 支付 1.25 美元，时薪为 15 美元。

在人类判断测试中，我们首先进行了资格审核轮次，以确保标注质量。在这一轮中，我们要求标注者完成一个包含 10 个示例（5 个无歧义、5 个有歧义，顺序随机）的单个人类智能任务。对于完成这一轮的每位标注者，我们通过统计其正确选择指称对象（或歧义标签）的频率来计算准确率，不考虑其置信度。排名前 100 的标注者获得了参与主要任务的资格。

在主要任务中，我们总共标注了 625 个句子，每个句子分配了 3 项标注任务。我们要求每位标注者最多完成 5 个 HIT，以便获得多样化的判断结果。与资格审核轮次类似，我们要求每位标注者标出指称对象（或歧义标签）并给出置信度。我们将标注分为 5 种选项：（名词 A，确定）、（名词 A，可能）、歧义、（名词 B，可能）和（名词 B，确定）。每种模板类型的人工判断标签通过计算五种选项中各自的标注比例进行汇总。主要任务的标注界面如图所示3。

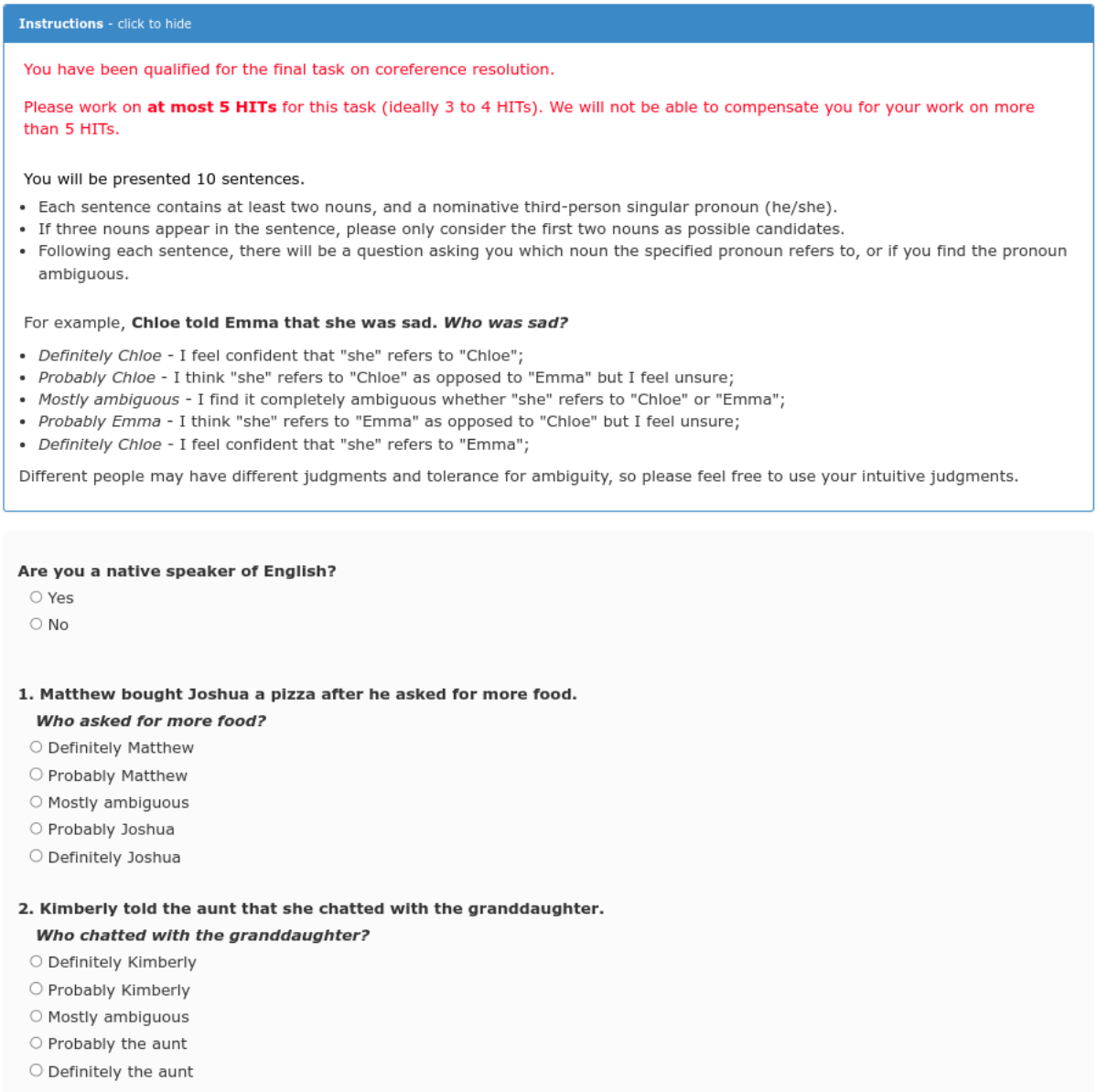

图 3：人工判断测试的标注界面，详见第 2 节。