用于评估机器翻译偏见的不确定性量化

Ieva Raminta Stalinait
University of Cambridge
irs38@cam.ac.uk

Julius Cheng
University of Cambridge
jncc3@cam.ac.uk

Andreas Vlachos
University of Cambridge
av308@cam.ac.uk

Abstract

在机器翻译（MT）中，当源语言句子包含一个性别没有明显标记的词素，但其目标语言的对应词需要明确性别时，模型必须从上下文和/或外部知识中推断出适当的性别。研究表明，机器翻译模型表现出偏见行为，甚至在与上下文信息冲突时仍依赖刻板印象。我们认为，除了在输入中明显显示时自信地使用正确性别进行翻译外，模型还应在性别模糊时保持不确定性。使用最近提出的语义不确定性指标，我们发现，即使在不明确的情况下翻译和性别准确性较高的模型，在模糊情况下也不一定表现出预期的不确定水平。同样，去偏在模糊和不模糊的翻译实例上具有独立的效果。[1]

1 介绍

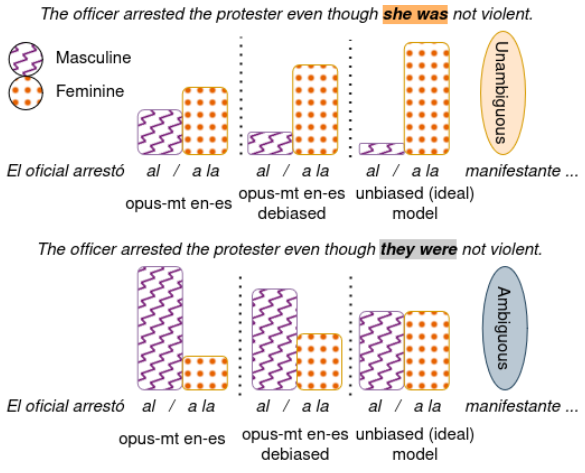

Figure 1: 在包含一个名词的句子的西班牙语翻译中，关于女性和男性限定词的概率（该名词是女性形式（称为“她”）或模糊的（“他们”）），由两个现有模型和理想的无偏模型预期归因。

语言本质上是模糊的，意义通常通过上下文来解决。然而，并不是所有的模糊性都是可以解决的 (van Deemter, 1998) 。当人类处理语言时，他们利用语言、认知和社会偏见来得出解释 (Cairns, 1973) 。尽管语言偏见可以简化认知处理，但有些也可能产生有害影响，例如加剧现有的社会不平等 (Beukeboom, 2013) 。NLP模型对许多人体偏见表现出敏感性 (Echterhoff et al., 2024) ，甚至夸大它们 (Dhamala et al., 2021) ，并引入额外的偏见 (Tjuatja et al., 2024) 。当输入模糊且无法解决时，倾向于单一输出必然依赖于偏见。因此，一个设计良好的模型应避免做出单一预测，而是请求澄清或生成多个替代输出。

大多数关于使用语言模型（LMs）进行机器翻译（MT）解码的研究评估单个预测，通常是通过束搜索为每个翻译实例生成一个预测。因此，关于不确定性量化（UQ）的大多数研究使用不确定性来预测通过束搜索恢复的翻译的质量 (Fomicheva et al., 2020; Cheng and Vlachos, 2024) 。关于 MT 偏差的先前工作集中于语言模型对抗黄金标准标签的性能 (Stanovsky et al., 2019) ，而关于 MT 歧义的先前工作同样集中于可以解析的案例 (Barua et al., 2024; Martelli et al., 2025) 。因此，大多数关于不确定性和歧义的工作假定每个实例都有一个单一正确的翻译。然而，对于模棱两可的源句，选择语言模型是否正确需要附加上下文才能确定，这方面的关注较少。这可以被视为一种不可减少的数据内在不确定性—即不确定性，称为 (Hora, 1996) 。根据 Baan et al. (2024) 的说法，语言模型中的概率质量分布既代表缺乏信心，也代表人类生成的多样性。我们采纳这种思路，并检验语言模型是否准确和公平地表示模棱两可源句的可能翻译范围。

在这项工作中，我们利用分布级的不确定性度量来评估由于输入的模糊性导致模型不应对其预测结果确定的情况。图 1 展示了一个句子的两个版本，其中名词“抗议者”的性别分别为明确的（顶部）和模糊的（底部），以及分配给该名词限定词的西班牙语翻译的不同概率。一个理想的模型应该在性别被代词明确化时，为女性限定词“la”分配更高的概率，而在性别模糊时，为男性和女性的翻译分配相同的概率。然而，最先进的机器翻译模型，包

[1]代码将在 https://anonymous.4open.science/r/uncertainty_bias_ambiguity-8A4C/ 发布

括去偏见的模型，往往会为明确的输入生成更均匀的概率分布，而为模糊的输入生成较不均匀的概率分布。这表明，模型概率受到了抗议和男性气质之间的刻板关联的影响，导致模型即使在没有性别偏好时也默认选择男性形式，并在有明确上下文线索的情况下以低信心选择女性形式。为了系统研究这一点，我们专注于将不对名词和动词进行性别标记的语言（英语）翻译成对名词和动词进行性别标记的语言（西班牙语、法语、乌克兰语和俄语）。我们使用了包含刻板性别角色的 WINOMT 数据集 (Stanovsky et al., 2019)，并通过额外的认知偏见线索（如隐含因果动词）的手动翻译和自动注释来扩展它。在某些情况下，性别可以从上下文中解析，而在其他情况下则不能。我们探索了语义不确定性指标 (Cheng and Vlachos, 2024; Farquhar et al., 2024) 的不同变体，用以量化翻译样本的语义多样性，发现这些指标有效地捕捉了由偏见触发导致的性别变异。我们根据既定的性别准确性指标验证了这些指标，并考虑了翻译准确性的影响。我们的主要发现是：1）刻板印象和语言偏见影响性别翻译，2）在明确情况下，偏见程度与整体模型翻译准确性对应，3）在模糊情况下，偏见程度与实例级别的翻译准确性对应，4）去偏效应因输入不明确性、翻译准确性和目标语言而异。

2 相关工作

研究人员已经解决了机器翻译中的各种偏见，包括算法偏见 (Vanmassenhove et al., 2021)，以及性别、数量和正式性偏见 (Měchura, 2022)。性别刻板印象不仅由语义内容触发，还包括说话方式 (Dawkins et al., 2024) 和人名 (Saunders and Olsen, 2023)。现有的性别翻译解决方案在明确 (Robinson et al., 2024) 和模糊 (Cho et al., 2019; Gonen and Webster, 2020; Vanmassenhove and Monti, 2021) 情况下依赖于工具或人工注释性别，从而限制了它们推广到其他类型歧义的能力。例如，Cho et al. (2019) 通过为具不明确代词的句子生成多种翻译来探索无法解决的歧义，然而这仅限于特定的句子类型。

在自然语言处理和机器学习研究中，已经提出了区分随机不确定性和认知不确定性的方法 (Hou et al., 2024)，然而它们并没有区分数据的随机性和数据的模糊性。有些研究关联了偏差与不确定性 (Sicilia et al., 2024; Kuzucu et al., 2025)，以及模糊性与不确定性之间的关系 (Kim, 2025; Cheng and Amiri, 2024)，然而，这些论文将不确定性解读为性能不佳的信号，将模糊性解读为低质量输入，这与我们对模糊性作为语言不可或缺特征的定义不同。问答领域的研究发现，检测模糊输入的最佳方法是量化模型输出样本中的重复性 (Cole et al., 2023)，并使用白盒指标如熵 (Yang et al., 2025)。

机器翻译中的不确定性量化（UQ）已被用作质量评估（QE）的代理工具。例如，Fomicheva et al. (2020) 使用机器翻译模型的不确定性来估计翻译质量，而不需要参考译文；而 Glushkova et al. (2021) 将相同的技术应用于质量评估模型本身的不确定性。其他方法则利用 UQ 来识别难以处理的实例，并通过应用课程学习 (Zhou et al., 2020)、语义增强 (Wei et al., 2020)、平衡多语言训练数据 (Wu et al., 2021) 或测试时适应 (Zhan et al., 2023) 来增强训练。Wang et al. (2024b) 考察了零样本翻译，并区分了模型不确定性和数据不确定性，然而他们对数据不确定性的关注更多在于嘈杂、低质量的训练数据，而非固有的歧义。认知科学研究已表明，基于熵的不确定性度量是人类翻译中歧义的合适测量工具 (Bangalore et al., 2016)，但这一见解尚未应用于机器翻译。据我们所知，过去没有基于 UQ 的机器翻译方法将歧义作为一种特殊的数据不确定性进行探索。

3 方法

我们提出了一种通过描述性别如何在预测分布中分配给名词的方式来量化神经机器翻译（NMT）模型中的性别偏见的方法。为此，我们基于最近提出的 UQ 指标，这些指标建立在经典的 Shannon 熵基础上，但考虑了来自模型的随机蒙特卡洛样本之间的相似性。我们首先对这些 UQ 方法提供一个简要概述。

设 $\mathcal{Y}$ 为一个随机变量，其值取自 NMT 模型 $p(y|x)$ 的预测分布。那么熵定义为：

$$\mathcal{H}(\mathcal{Y}) = \mathop{\mathbb{E}}_{y\sim\mathcal{Y}}[I(y)],$$

，其中 I 是 y 的意外性。在经典的香农熵中，$I = -\log p(y)$，但我们考虑的不确定性量化方法在其意外性定义中有所不同。

语义熵（SE；**?**）识别元素之间的语义等价性，并根据文本蕴涵模型将它们聚集在一起，将每个 y 映射到一个簇 c。在我们的实现中，我们使用经过多语言 mDeberta 模型 (He et al., 2021)，由 Laurer et al. (2022) 在自然语言推理（NLI）任务上进行微调。然后，意外率是元素位于 c 中的负对数概率：

$$I_{\text{SE}}(y) = -\log \mathop{\mathbb{E}}_{y'\sim\mathcal{Y}} 1\left[y' \in c\right].$$

相似性敏感的 Shannon 熵（S3E；Ricotta and Szeidl, 2006; **?**）将 y 的意外性设置为其与所有

其他输出的预期相似度的负对数：

$$I_{\text{S3E}}(y) = -\log \mathbb{E}_{y' \sim \mathcal{Y}} \left[\mathcal{S}(y, y') \right],$$

这里 $\mathcal{S}$ 是一个相似度函数，满足 $\mathcal{S}(y, y') \in [0, 1]$ 和 $\mathcal{S}(y, y') = 1$ 条件，如果 $y = y'$ 。遵循 Cheng and Vlachos (2024) 的做法，我们使用由多语言 E5 文本嵌入模型 (Wang et al., 2024a) 生成的 y 和 y' 句子嵌入的余弦相似度。

我们还定义了性别熵（GE），其计算方式类似于 SE，但基于翻译后的焦点名词的性别类别对元素进行聚类。为了确定性别类别，我们使用 Spacy [2] 和 pymorphy2 (Korobov, 2015) 形态解析器。

必须从 $p(\cdot|x)$ 中通过随机采样近似 SE 、S3E 和 GE，我们执行的方法是 ϵ -采样 (Hewitt et al., 2022)，每个源句子抽取 128 个样本。关于这些 UQ 方法的更多细节可以在附录 A 中找到。

我们的性别偏见指标基于这些不确定性量化（UQ）方法给出的意外性和熵。第一个理想条件是，对于性别明确的源句子，一个无偏见的模型应当使正确性别词形变化的翻译相比于错误词形变化具有更低的意外性。因此，无偏见的模型应该最小化相对意外性，定义为：

$$\Delta I = \frac{I(y_{\text{correct}}) - I(y_{\text{incorrect}})}{\frac{1}{2}(I(y_{\text{correct}}) + I(y_{\text{incorrect}}))}.$$

第二个期望是无偏模型的熵不应受到偏见提示的影响。因此，我们定义了归一化熵，它将源码句子 x 的 $\mathcal{H}$ 与其对比集 $\mathcal{G}_x$ 的平均熵进行比较。$\mathcal{G}_x$ 是一组最小化不同的句子，除了代词（例如，“她”，“他”，“他们”）之外，它们与 x 相同，包括 x 本身。表 1 中的三个句子组成了一个 $\mathcal{G}_x$ 。正式地：

$$\text{norm-}\mathcal{H}(x) = \frac{\mathcal{H}(\mathcal{Y}_x)}{\frac{1}{|\mathcal{G}_x|} \sum_{x' \in \mathcal{G}_x} \mathcal{H}(\mathcal{Y}_{x'})},$$

这个公式通过在对比集中保持其他所有词汇、句法和语义内容不变，专门隔离了归于性别的 $\mathcal{H}(\mathcal{Y}_x)$ 的变化。

第三个期望是，与不含歧义的输入相比，模型对于性别模棱两可的输入应表现出更高的不确定性，不考虑输入中的所有偏见。因此，应最小化相对熵，其定义为：

$$\Delta\mathcal{H} = \frac{\mathcal{H}(\mathcal{Y}_{\text{unambiguous}}) - \mathcal{H}(\mathcal{Y}_{\text{ambiguous}})}{\frac{1}{2}\left(\mathcal{H}(\mathcal{Y}_{\text{unambiguous}}) + \mathcal{H}\left(\mathcal{Y}_{\text{ambiguous}}\right)\right)}.$$

[2] https://spacy.io/

4 实验设置

为了测试我们提出的偏见指标，我们需要一个包含关于源语句中性别模糊性和刻板印象信息的机器翻译数据集。我们使用 WINOMT (Stanovsky et al., 2019)，其中包括对 1,584 个句子的最小对的注释，这些句子中使用的阳性、阴性或中性代词指代典型或非典型性别角色。表格 1 中可以看到数据集中的三句话例子。基于上下文信息（代词 M & F），“mechanic”一词的性别在前两句话中是明确的，但由于中性（N）代词，第三句依然模糊。关于机械师通常是男性而非女性的刻板印象（第 3 列）要么与消歧上下文（第 2 列）相矛盾（第 1 行），要么一致（第 2 行）。

该数据集还包含在其发布版本中未经明确标注的其他语言现象。因此，我们使用 Spacy 提供的句法解析自动标注了额外的语言偏见提示，即主语、新近性、隐含因果关系和人名。我们将额外的标注公开以便于复现。

关于主语偏差，依据关于人类对连贯性偏见的文献 (Nieuwland and Van Berkum, 2006)，我们假设模型可能会假定主句的主语（通常是焦点名词）与补语从句的主语（通常是代词）之间有照应关系。在表格 1 的例子中，“机械师”是主语，因此主语偏差启动了一种解释，即主语的性别与后续代词的性别一致（第一行是男性，第二行是女性）。此外，人名已被证明对代词解析有很大的影响 (Saunders and Olsen, 2023)。为了评估人名对性别翻译的影响，我们通过添加与代词性别匹配的常见女性和男性名字来增强数据集，这些名字是根据它们在跨语言中的熟悉程度选择的（见附录 B）。例如，当翻译成法语时，表格 1 中的第二句将读为“机械师 Anne 打电话通知某人她已完成修理。”

新近偏倚会导致将最近的名词短语的性别归因于后续的代词 (Gautam et al., 2024)。然而，在我们的例子中，最近的名词短语是“someone”，因此焦点名词不受影响。隐含因果（IC）偏倚在人类和语言模型中诱发一种期望，即当 IC 动词后跟因果连词时，后续的代词将与 IC 动词的主语或宾语共同指代，具体取决于因果推理 (Kementchedjhieva et al., 2021)。例如，一个带有动词“call”的从句，当后面跟随一个以“because”开始的解释时，预期会提到打电话的人而不是接电话的人。我们使用一个 IC 动词语料库 (Garnham et al., 2021) 进行标注，发现大约 10 % 的数据集包含 IC 结构。表 1 中的句子不包含因果补语，因此不适用 IC 偏倚。

目标语言 我们选择了通过名词和形容词上的形态标记表达性别的目标语言，有时也包括

Sentence	Pronoun	Stereotype	Subject	Recency	IC
The mechanic called to inform someone that he had completed the repair.	M	M	M	N	N
The mechanic called to inform someone that she had completed the repair.	F	M	F	N	N
The mechanic called to inform someone that they had completed the repair.	N	M	N	N	N

Table 1: WINOMT (Stanovsky et al., 2019) 示例以及偏见提示的附加注释。

动词，即西班牙语、法语、乌克兰语和俄语。这些目标语言在自然语言处理研究中的表现各不相同。俄语语言的一个相关特性是，对于一些描述职业的名词，即使存在女性化形式，使用时也可能被认为是带有贬义的 (Komova, 2024) 。例如，‘ ’ 是 ‘医生’ 的阳性词语，而其对应的女性化形式 ‘ ’ 被认为是粗鲁的，因此即使已知医生是女性，仍然使用 ‘ ’。这导致了使用阳性名词与阴性动词形式搭配的结构，或者句中始终使用阳性标记。为了考虑这一点，我们还包括了对遵循这种（缺乏）性别标记的职业分类为 WINOMT ，使用的是来自 Komova (2024) 的数据。

人工翻译 WINOMT 不包含目标翻译，因此无法直接评估机器翻译模型的准确性。为了解决这个问题，我们聘请了专业译者将一组 100 个 WINOMT 句子翻译成法语、西班牙语、乌克兰语和俄语。每个句子都被翻译两次，分别以阴性和阳性变体关注名词。他们还根据给定上下文中的性别翻译将翻译标注为 “正确” 或 “不正确”。例如，当将英语句子 “农夫从作家那里买了一本书并付给她” 翻译成法语时，其中 “作家” 是关注名词，应该将阴性的 “l’auteure” 标记为正确，而 “l’auteur” 被视为不正确。在模棱两可的情况下，比如上述句子中的代词是 “他们”，则两种性别的翻译都被视为正确。附录 C 提供了翻译指南和细节，附录 D 讨论了人工标注的质量。我们将翻译和正确性标注向公众发布，以促进进一步的研究。

我们尝试了两种常用的翻译模型，即 OPUS-MT (Tiedemann and Thottingal, 2020) 和 M2M100 (Fan et al., 2021) 。OPUS-MT 模型是在开放获取的平行语料库上训练的 NMT 模型。M2M100 是一种多对多的多语言翻译模型，可以直接在 100 种语言中的任意一对之间进行翻译。为了检查消除偏见措施在第 3 节中提到的三个愿望方面的效果，我们对 OPUS-MT 模型应用了来自 Iluz et al. (2024) 的硬偏置消除方法，该方法已被证明能够在 WINOMT 数据集 (Stanovsky et al., 2019) 上降低偏置评分，同时保持翻译质量。硬偏置消除方法在表示空间中中和了偏置词，使中性词不与特定性别相关联 (Bolukbasi et al., 2016) 。我们采用了来自 Iluz et al. (2024) 的最有效的偏置消除方法，该方法对编码器端的单标识符职业词进行了偏置消除。所有模型的性能见附录 E 。

5 研究问题

为了验证 UQ 度量在偏差评估中的应用，我们将其分数与已建立的性别准确性度量进行比较。性别准确性使用第 3 节中描述的形态解析器来确定翻译中的焦点名词性别。由于其依赖于黄金标准的参考，它仅适用于明确无误的项目，不适合有多个有效性别实现的情况。因此，我们将此实验限定为明确无误的项目。我们根据模型的 ΔI 分数对所有模型进行排名，并使用 Kendall’s τ 和 Pearson’s r 将该排名与基于性别准确性的排名进行比较。为了确定是否需要基于采样的度量，我们还测试一个简单的 $\Delta LogProb$ 值，作为 ΔI 的替代方案，该值用于比较分配给正确和错误实例的对数概率。

为了评估模型的偏差，我们评估偏差线索如何影响翻译中性别标记的多样性。具体来说，我们进行方差分析（ANOVA ）以检查偏差线索（自变量）对归一化熵测量 norm- $\mathcal{H}$ （S3E ）、norm- $\mathcal{H}$ （SE ）和 norm- $\mathcal{H}$ （GE ）（因变量）的影响，并使用 T 检验判断显著性。此分析包含了以前在 WINOMT 中未探索过的语言偏差。

为了评估在尚未被探索的模糊环境中的模型偏差，我们通过它们的 $\Delta\mathcal{H}$ 分数来比较模型。为了将由模糊性引起的不确定性与模型性能不佳导致的不确定性分开，我们分析 $\Delta\mathcal{H}$ 分数与通过 COMET 度量 (Rei et al., 2022) 测量的翻译质量之间的关系。由于 WINOMT 数据集不包含标准的目标翻译，我们使用第 4 节中描述的 100 个专业注释项目和 WMT 测试集 (包含目标语言 (Callison-Burch et al., 2012; Bojar et al., 2013, 2014; Koehn et al., 2023; Haddow et al., 2024)) 进行翻译质量评估。

6 结果

本节展示了使用语义不确定性度量进行偏差评估的结果。

根据语义意外值和性别准确率的模型排名相关。 第一次实验的结果如表 2 所示。我们发现虽然 ΔLog prob 与性别准确性排名没有相关性，但 ΔI （S3E ）与性别准确性呈现统计学上显著的负相关（肯德尔的 $\tau = -0.58$ ，斯皮尔曼的 $\rho = -0.78$ ；见附录 K ）。我们认

Lang.	Model	Gender Acc	ΔLog prob	ΔI (S3E)	COMET
ES	OPUS-MT	67.95	0.00	-0.10	84.90
	deb-OPUS-MT	68.13	0.00	-0.13	84.86
	M2M100	70.77	0.00	-0.13	72.05
FR	OPUS-MT	64.27	0.01	-0.04	83.56
	deb-OPUS-MT	64.79	0.01	-0.08	83.55
	M2M100	61.66	0.01	-0.07	73.06
UK	OPUS-MT	45.34	0.00	-0.03	70.79
	deb-OPUS-MT	46.12	0.00	-0.03	70.79
	M2M100	47.76	0.00	-0.02	52.85
RU	OPUS-MT	48.57	0.00	0.00	79.37
	deb-OPUS-MT	48.42	0.00	-0.03	79.36
	M2M100	48.49	0.00	-0.03	58.62

Table 2: 在明确的实例上，性别准确性、Δ 对数概率和 ΔI （S3E），以及 WMT 测试集上的 COMET 分数（详见附录 E）。

为该指标在区分正确与不正确性别翻译方面的有效性，归因于其能够通过嵌入表示捕捉超越名词形态的性别信息，包括动词词尾变化和一致性。S3E 的灵活性使其能够编码 GE 无法体现的细微差别。例如，当将一个包含女性代词的句子翻译成俄语时，OPUS-MT 生成句子时可能会用一个阳性名词和一个女性或男性词尾的动词（例如，“ （有限元法）/ （雄性）”“快递员表示感谢”）。这反映在 S3E 的 $\mathcal{H}$ 得分（0.65）高于 GE （0.00），因为仅有动词词尾变化的变体被 S3E 捕捉到。

ΔI （S3E）与性别准确性之间的强负相关因此验证了我们所提出的用于评估机器翻译偏见的指标的核心组成部分。

此外，模型按照其整体性能（由 COMET 分数表示）的排名与基于性别准确性和不明确实例的 ΔI 排名部分一致。这表明在这些实例中，表现更好的模型往往偏见较小（所有排名列在附录 J 中）。

语义熵分数随偏差信号的变化而变化。 表格 3 中展示的第二个实验结果显示，数据中的大多数偏见线索对范数- $\mathcal{H}$（S3E）的方差有显著影响，这表明受测模型表现出各种社会和语言偏见。[3] 这些结果证实了之前的研究发现。Names 列中的高绝对系数值表明，即使在出现消歧代词时，人名对性别翻译也有影响。这与 Saunders and Olsen (2023) 的结论一致，表明代词和姓名都会导致性别偏见，且通常不足以实现完全消歧。其次，部分俄语名词在任何语境中都有默认男性语法性别的事实反映在含有这些名词的句子的性别多样性显著下降（Default M 列中的负系数表明较低的范数- $\mathcal{H}$）。第三，我们观察到男性偏向通常会降低范数- $\mathcal{H}$ （M 列中的负系数），而女性偏向则倾向于增加它（F 列中的正系数）。这表明模型在翻译中通常默认为男性翻译，输出在男性偏见下变得更相似，而在女性偏见下则更为多样化。该发现与 Kuzucu et al. (2025) 的结论一致，表明模型的不确定性对于少数群体通常更高。

翻译准确性在不同分析层面上对偏差-熵关系的影响各异。 在第一个实验中验证了 S3E 度量后，我们研究了 $\Delta\mathcal{H}$（S3E）的结果作为偏见度量。表格 4 表明一些模型（OPUS-MT -UK, deb- OPUS-MT -UK, M2M100 -UK, OPUS-MT -RU, M2M100 -RU）表现出所需的负 $\Delta\mathcal{H}$。令人惊讶的是，这一结果表明在处理含糊实例时，与明确实例相反，那些在翻译准确性上表现更好的模型（即西班牙语和法语模型）通常并不表现出较少的性别偏见（根据所有度量的模型排名在附录 J 中提供）。这一发现反映了表格 3 中“模糊性”列的结果，其中乌克兰语和俄语的 norm- $\mathcal{H}$ 增加（正系数），而西班牙语和法语则没有。对于无偏模型，预期模糊项目的 norm- $\mathcal{H}$ 较高（或负的 $\Delta\mathcal{H}$）。

相比之下，我们观察到西班牙语和法语模型中去偏的预期效果（在表格 4 中，deb- OPUS-MT 的 $\Delta\mathcal{H}$ 较低），这表明整体性能更好的模型更容易受到去偏的影响。去偏的影响也体现在表格 3 中，因为去偏后的模型在各语言中的效果大多较小（系数的绝对值较低），这证实了去偏至少是部分有效的。

在图 2 中，结果按 COMET 分数组进行分组，以便在实例级别进行更细粒度的分析。对于具有负 $\Delta\mathcal{H}$ （S3E）分数的模型（乌克兰语和俄语），$\Delta\mathcal{H}$ 通常在最高准确度的翻译中最为明显（例如，M2M100 -RU 在 Bin 3 中的模糊分数明显高于 B1）。尽管去偏没有减少乌克兰语的整体 $\Delta\mathcal{H}$ 分数（见表 4），但它在高质量翻译中带来了最大的提升：在 B3 分组中，deb-OPUS-MT -UK 的模糊 $\mathcal{H}$ 分数显著高于原始模型的分数。这一改进进一步得到乌克兰语中阳性焦点名词屈折变化显著下降 8.41 % 的支持，而其他语言的降幅为 0.88-2.49 %。我们假设这是由于乌克兰语的训练数据有限，可能导致模型整体表现较差但在高质量输出中对去偏更为敏感。翻译准确度与在歧义下偏差之间的关系似乎根据分析级别而有所不同：在不同模型之间，较高的准确度并不意味着在歧义实例上偏差较低，而在同一模型内，准确度较高的实例往往表现出较低的偏差。

定性分析 对表格 1 的示例进行的定性分析，与表格 5 中的对应 $\mathcal{H}$ 值一起提供，证实了

[3] 与 norm- $\mathcal{H}$（SE）和 norm- $\mathcal{H}$（GE）相比，norm- $\mathcal{H}$（S3E）对偏差线索表现出最强的敏感性。我们也对未归一化的 $\mathcal{H}$ 分数进行了实验，其结果在不同指标、偏差类型和模型之间的可比性较差。完整的结果在附录 F 中呈现。对于 norm- $\mathcal{H}$（S3E）观察到的所有趋势，在 norm- $\mathcal{H}$（SE）和 norm- $\mathcal{H}$（GE）以及未归一化的 $\mathcal{H}$ 中也存在。

Lang.	Model	Names	Recency		Implicit Causality				Stereotype				Subject		Pronoun				Default M		Ambiguity
			F	M	SF	SM	OF	OM	SF	SM	OF	OM	F	M	SF	SM	OF	OM	S	O	
es	OPUS-MT	0.41	0.41	-0.05	0.25	-0.19	0.24	-0.33	0.06	0.10	0.13	0.17	0.38	-0.21	0.24	-0.31	0.29	-0.13	N/A	N/A	-0.18
	deb-OPUS-MT	-0.05	0.30	-0.11	0.27	-0.20	0.16	-0.38	0.05	0.14	0.08	0.05	0.49	-0.14	0.39	-0.24	0.14	-0.21	N/A	N/A	-0.10
	M2M100	0.14	0.33	-0.11	0.29	-0.42	0.28	-0.25	0.18	0.27	0.12	0.07	0.51	-0.04	0.52	-0.08	0.04	-0.35	N/A	N/A	-0.11
fr	OPUS-MT	0.54	0.42	0.16	0.05	-0.34	0.06	-0.24	0.43	0.45	0.26	0.21	0.16	-0.14	0.16	-0.17	0.20	-0.04	N/A	N/A	-0.29
	deb-OPUS-MT	0.19	0.22	0.02	0.05	-0.25	0.17	-0.11	0.23	0.32	0.23	0.13	0.24	-0.03	0.31	-0.08	0.01	-0.14	N/A	N/A	-0.12
	M2M100	-0.12	0.11	-0.23	0.50	0.09	0.49	0.03	-0.03	0.11	0.21	0.08	0.70	0.31	0.47	0.05	-0.08	-0.40	N/A	N/A	0.06
uk	OPUS-MT	-0.27	0.00	-0.11	0.26	-0.02	0.19	0.13	-0.11	0.17	0.03	-0.14	0.27	0.21	0.22	0.13	-0.11	-0.23	N/A	N/A	0.06
	deb-OPUS-MT	-0.43	-0.06	-0.12	0.41	0.00	0.18	0.07	-0.24	0.01	0.01	-0.17	0.23	0.17	0.16	0.10	-0.11	-0.17	N/A	N/A	0.09
	M2M100	0.08	-0.04	-0.18	0.27	0.02	0.33	0.18	0.19	0.37	-0.03	-0.17	0.53	0.34	0.50	0.26	-0.30	-0.42	N/A	N/A	0.11
ru	OPUS-MT	0.01	-0.28	-0.41	0.00	-0.12	0.08	-0.10	-0.19	-0.20	-0.41	-0.39	0.07	-0.03	0.14	0.03	-0.10	-0.24	-0.40	0.04	0.35
	deb-OPUS-MT	0.23	-0.10	-0.17	0.05	0.03	0.05	-0.03	-0.10	-0.04	-0.11	-0.14	0.09	0.04	-0.13	0.03	-0.08	-0.13	-0.26	0.00	0.13
	M2M100	-0.74	-0.16	-0.21	0.10	-0.04	-0.12	-0.25	0.12	0.12	-0.20	-0.18	0.06	-0.02	0.08	0.01	-0.07	-0.11	-0.20	-0.22	0.18

Table 3: ANOVA 结果：偏倚提示（F 为女性化，M 为男性化，S 为主体，O 为宾体）对标准 $\mathcal{H}$（S3E）的单一影响。数值对应于效应系数（相对于参考组的偏差）。粗体字表示具有统计显著性（$p < 0.05$）。数值的符号表示该变量的存在是增加（正）还是减少（负）包含该变量值的组的平均 $\mathcal{H}$。除了：名字的“无名称”、默认 M 的“无默认”、歧义的“无二义性”之外，参考组为所有列的 N。

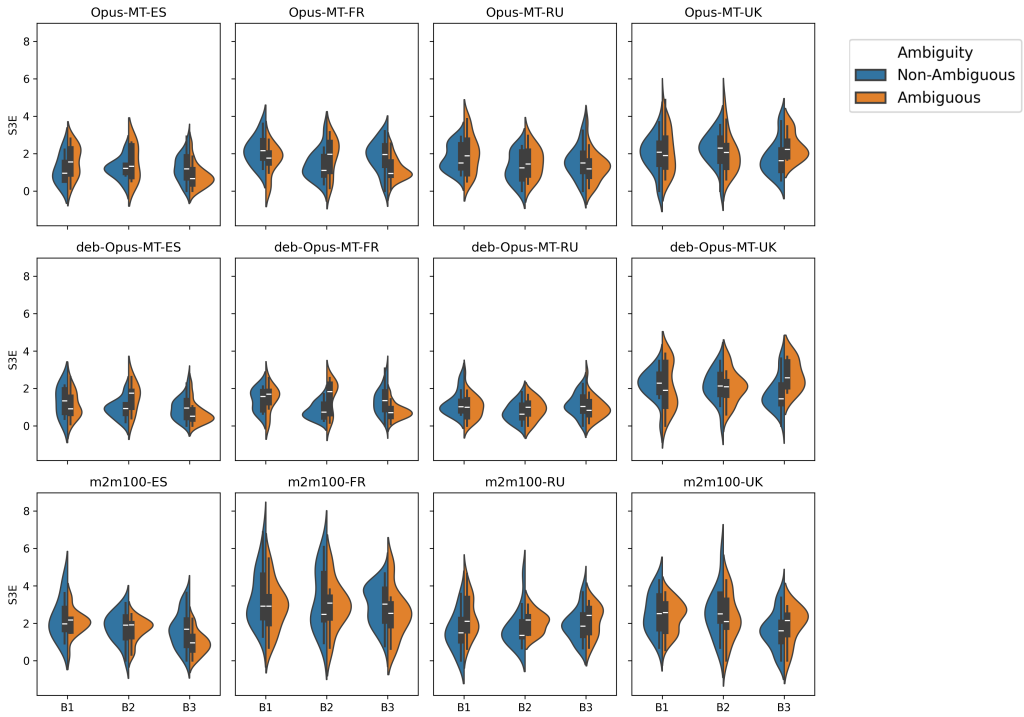

Figure 2: 小提琴图展示了模糊和非模糊输入上分组的 COMET 分数和 $\mathcal{H}$（S3E）。从左到右分别是低、中和高 COMET 分数，使用人工翻译和模糊项目的多参考进行评估。

图 2 中的定量发现。在翻译第 1 行的句子时，对于所有目标语言，OPUS-MT 模型始终仅产生焦点名词的阳性变体（‘El mecánico’、‘Le mécanicien’、‘ ’和‘ ’）。在反刻板印象的情况下（第 2 行），除俄语外的所有语言都包含阳性和阴性形式（‘La mecánica’、‘La mécanicienne’ 和‘ ’），这表明这些模型即使在语境中的指代对象明显是阴性时也对男性刻板印象敏感。对于俄语，即使在语境明确为阴性的情况下，这些模型也无法生成任何阴性结构。这种差异在 $\mathcal{H}$（S3E）分数中很明显，对于前三种语言，第 2 行的分数高于第 1 行。此外，在模糊情况下（第 3 行），所有 OPUS-MT 模型仅产生阳性名词，无论语言如何。因此，$\mathcal{H}$（S3E）分数在明确情况下（第 1 和第 2 行的平均值）通常高于模糊情况（第 3 行），除了俄语，$\mathcal{H}$ 在所有情况下始终较低。这些观察结果与对有偏模型的预期一致：当代词模糊时，它们默认为刻板的性别呈现，有时即使语境明显暗示反刻板印象的解释也是如此。

Lang.	Model	Unamb	Amb	$\Delta\mathcal{H}$
ES	OPUS-MT	1.23	1.12	0.09
	deb-OPUS-MT	0.97	0.89	0.08
	M2M100	1.79	1.45	0.19
FR	OPUS-MT	1.79	1.43	0.20
	deb-OPUS-MT	1.21	1.08	0.11
	M2M100	3.22	2.78	0.14
UK	OPUS-MT	1.96	2.16	-0.10
	deb-OPUS-MT	1.98	2.15	-0.09
	M2M100	2.05	2.28	-0.11
RU	OPUS-MT	1.56	1.68	-0.08
	deb-OPUS-MT	1.05	0.97	0.08
	M2M100	1.83	2.29	-0.25

Table 4: 不明确和明确的 $\mathcal{H}$ （S3E）

定性分析还揭示了去偏的有趣效果。在第 2 行反刻板印象的情况下，去偏增加了西班牙语中女性结构的生成数量（从 43/128 增加到 55/128，对应于 $\mathcal{H}$ 的轻微增加，因为女性形式仍然是少数）和乌克兰语（从 72/128 增加到 128/128，反映在 $\mathcal{H}$ 的减少中）。对于法语或俄语，没有观察到显著变化，与稳定的 $\mathcal{H}$ （S3E）评分一致。对于模糊代词（第 3 行），去偏模型在西班牙语、法语和俄语中继续仅生成“mechanic”的男性变体，且 $\mathcal{H}$ 基本保持不变。相比之下，乌克兰语中的所有去偏模型输出都包含了名词的女性翻译，对应于第 3 行中 $\mathcal{H}$ 从 OPUS-MT 到去偏 OPUS-MT 的减少。这种模式表明，当去偏在模糊情境中导致女性形态的过度生成时，我们提出的指标将此标记为偏差增加（正 $\Delta\mathcal{H}$），这表明此类变化未被视为改进。

7 结论

在这项工作中，我们应用分布级别的不确定性量化（UQ）来评估机器翻译（MT）模型中的偏差。这种方法能够补充性别准确性，特别是在性别准确性不适用的情况下。具体而言，它能够捕捉更为微妙的性别偏见表现，这些偏见在模型对性别模糊情境显示偏好时出现。我们总体的贡献在于首次将 UQ 用作 MT 中的偏差度量，这种方法 1）不依赖性别参考，2）是通用的并能捕捉多种类型的偏见，3）通过既有的性别准确性度量进行了验证，以及 4）提供了在以前未被研究的性别模糊情境下有关偏见行为的新见解。未来的工作将把我们提出的偏差评估方法扩展到除翻译以外的任务中。

8

局限性

本研究仅限于罗曼语族和斯拉夫语族，不包括许多其他以多种方式标记性别和表达刻板印象的语言家族。虽然我们试图通过为不同目标语言包括不同的名字、考虑俄语中特定的男性专有名词、使用特定语言的词汇去偏化等方法来考虑语言差异，但仍然有一些语言特性未被考虑，例如职业刻板印象在英语中定义，可能在不同地区应用有所不同。最后，我们的工作仅限于两个语法性别，并将“they”作为一个可以指代任何性别的中性代词，而我们未专门研究其作为指代非二元性别者的代词的解释。进一步的研究方向包括将不确定性量化（UQ）应用于歧义检测，这可以通过形态双重表达实现更具包容性的性别翻译，其中包含男性和女性词素以实现性别中立。未来的工作应解决这些方向。

9

伦理声明

本研究中使用的模型与所有机器学习模型一样，可能存在偏见并犯错，包括在性别归因上。我们的贡献旨在通过更严格的评估指标，专门解决男性和女性性别刻板印象问题，以避免性别偏见的延续。

10

致谢 Ieva Raminta Stalinait 得到了华为的支持。Andreas Vlachos 得到了 ERC 资助 AVeriTeC (GA 865958) 的支持。Julius Cheng 得到 Imminent 的研究资助支持。数据集的翻译由同一个 Imminent 研究资助资助，我们想感谢 Andrii Nikolaiev、Mykhailo Khokhlovych、Sarah Mokadem 和 Daniel Carter 在翻译工作中的贡献。

References

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.

Srinivas Bangalore, Bergljot Behrens, Michael Carl, Maheshwar Ghankot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm. 2016. *Syntactic Variance and Priming Effects in Translation*, pages 211–238. Springer International Publishing, Cham.

Josh Barua, Sanjay Subramanian, Kayo Yin, and Alane Suhr. 2024. Using language models to disambiguate lexical choices in translation. In *Proceedings of*

Sentence	ES → deb		FR → deb		UK → deb		RU → deb	
The mechanic called to inform someone that he had completed the repair.	0.75	0.82	0.00	0.00	2.41	1.57	0.33	0.31
The mechanic called to inform someone that she had completed the repair.	1.64	1.85	0.53	0.56	2.75	2.05	0.00	0.00
The mechanic called to inform someone that they had completed the repair.	0.74	0.87	0.00	0.00	2.38	2.03	0.37	0.37

Table 5: 对于 OPUS-MT （左）和 deb- OPUS-MT （右）模型，具有 $\mathcal{H}$ （S3E ）值的 WINOMT 示例。

the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4837–4848, Miami, Florida, USA. Association for Computational Linguistics.

Camiel J Beukeboom. 2013. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication*, pages 313–330.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia, editors. 2013. *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors. 2014. *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Helen S Cairns. 1973. Effects of bias on processing and reprocessing of lexically ambiguous sentences. *Journal of Experimental Psychology*, 97(3):337.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia, editors. 2012. *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada.

Jiali Cheng and Hadi Amiri. 2024. FairFlow: Mitigating dataset biases through undecided learning for natural language understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21960–21975, Miami, Florida, USA. Association for Computational Linguistics.

Julius Cheng and Andreas Vlachos. 2024. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128, St. Julian's, Malta. Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Hillary Dawkins, Isar Nejadgholi, and Chi-Kiu Lo. 2024. WMT24 test suite: Gender resolution in speaker-listener dialogue roles. In *Proceedings of the Ninth Conference on Machine Translation*, pages 307–326, Miami, Florida, USA. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the*

Association for Computational Linguistics, 8:539–555.

Alan Garnham, Svenja Vorthmann, and Karolina Kaplanova. 2021. Implicit consequentiality bias in english: A corpus of 300+ verbs. *Behavior Research Methods*, 53:1530–1550.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors. 2024. *Proceedings of the Ninth Conference on Machine Translation*. Association for Computational Linguistics, Miami, Florida, USA.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephen C. Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2):217–223. Treatment of Aleatory and Epistemic Uncertainty.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Bar Iluz, Yanai Elazar, Asaf Yehudai, and Gabriel Stanovsky. 2024. Applying intrinsic debiasing on downstream tasks: Challenges and considerations for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14914–14921, Miami, Florida, USA. Association for Computational Linguistics.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised Mary because _he_? implicit causality bias and its interaction with explicit cues in LMs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4859–4871, Online. Association for Computational Linguistics.

Hazel H. Kim. 2025. How ambiguous are the rationales for natural language reasoning? a simple approach to handling rationale uncertainty. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10047–10053, Abu Dhabi, UAE. Association for Computational Linguistics.

Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors. 2023. *Proceedings of the Eighth Conference on Machine Translation*. Association for Computational Linguistics, Singapore.

Liliana Komova. 2024. *Gender, Language and Perception: Linguistic Inclusivity in Russian*. Ph.D. thesis, Università Ca'Foscari Venezia.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts*, pages 320–332, Cham. Springer International Publishing.

Selim Kuzucu, Jiaee Cheong, Hatice Gunes, and Sinan Kalkan. 2025. Uncertainty as a fairness measure. *Journal of Artificial Intelligence Research*, 81.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying –Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Federico Martelli, Stefano Perrella, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2025. Dibimt: A gold evaluation benchmark for studying lexical ambiguity in machine translation. *Computational Linguistics*, pages 1–71.

Michal Měchura. 2022. A taxonomy of bias-causing ambiguities in machine translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington. Association for Computational Linguistics.

Mante S. Nieuwland and Jos J.A. Van Berkum. 2006. Individual differences and contextual bias in pronoun resolution: Evidence from erps. *Brain Research*, 1118(1):155–167.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: Bridging the gap between the shannon entropy and rao's quadratic index. *Theoretical Population Biology*, 70(3):237–243.

Kevin Robinson, Sneha Kudugunta, Romina Stella, Sunipa Dev, and Jasmijn Bastings. 2024. MiTTenS: A dataset for evaluating gender mistranslation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4115–4124, Miami, Florida, USA. Association for Computational Linguistics.

Danielle Saunders and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93, Tampere, Finland. European Association for Machine Translation.

Anthony Sicilia, Mert Inan, and Malihe Alikhani. 2024. Accounting for sycophancy in language model uncertainty estimation. *arXiv preprint arXiv:2410.14746*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Kees van Deemter. 1998. Ambiguity and idiosyncratic interpretation. *Journal of Semantics*, 15(1):5–36.

Eva Vanmassenhove and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wenxuan Wang, Wenxiang Jiao, Shuo Wang, Zhaopeng Tu, and Michael R. Lyu. 2024b. Understanding and mitigating the uncertainty in zero-shot translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:4894–4904.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Luxi Xing, and Weihua Luo. 2020. Uncertainty-aware semantic augmentation for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2724–2735, Online. Association for Computational Linguistics.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7291–7305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863, Albuquerque, New Mexico. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao, and Min Zhang. 2023. Test-time adaptation for machine translation evaluation by uncertainty minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–820, Toronto, Canada. Association for Computational Linguistics.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.

A 进一步的不确定性量化讨论

Farquhar et al. (2024) 不定义每个元素的惊讶量；原始定义计算簇的香农熵。$\mathcal{Y}$ 被映射到簇 $\mathcal{C}$，以及 SE 是：

因此，

$$\mathcal{H}_{\text{SE}}(\mathcal{C}) = - \mathop{\mathbb{E}}_{c \sim \mathcal{C}} \log p(c|x).$$

意外性是针对簇而不是元素定义的，但很容易证明，我们的每元素意外性获得了与原始定义等效的熵。

Cheng and Vlachos (2024) 引入了一个超参数 α，该参数作为指数应用于相似性函数。这被用来重新调整 $\mathcal{S}$，以在基准测试中获得更有利的性能。我们调整 α，以实现 S3E 与形态分析器分配给相关名词的性别标签的熵之间的最高相关性。通过这种方式，我们希望 $\mathcal{H}$ (S3E) 能够捕捉性别变化，且更高的 $\mathcal{H}$ (S3E) 值表明性别词素的多样性更大。

我们还尝试了其他替代的相似性度量方式，比如欧几里得距离、切比雪夫距离、曼哈顿距离、闵可夫斯基距离等，但没有发现显著差异，因此在论文中使用了最简单的余弦相似性。

B 名称的注释

表 6 展示了用于扩展 WINOMT 数据集的名称，以包含法语、西班牙语、乌克兰语和俄语中的男性和女性常用名。

	Spanish	French	Ukrainian	Russian
F	Carla	Anne	Anna	Anna
M	Gabriel	Victor	Ivan	Ivan

Table 6: 用于扩展 WINOMT 数据集以适应不同目标语言的名称。

C 翻译指南和翻译招聘详情

根据要求，我们聘请了专业翻译人员，并按照每个单词 0.2 英镑支付报酬。翻译指南如下：

请将以下句子翻译成 [法语 /西班牙语 /乌克兰语 /俄语]。请为每个句子提供两个翻译，针对高亮显示的名词使用不同的性别变形——一个为阳性，一个为阴性。请在可能的情况下保持代词与原文一致。请使用作为母语者的您自然感到合适的翻译，而不是使用规定的形式 [4]。请注意这两种翻译在名词的性别与语境信息匹配方面是否都可接受，如果不接受，那么哪个是正确的：

- 开发人员与设计师争论，因为她不喜欢这个设计。
- 本科生向管理员提交了他们组织的请愿书。

[4]由于译者询问关于使用正式或主要方言以及语言政策变化对某些词汇翻译的影响，增加了这一说明。

D 人工标注质量

由专业翻译员进行的正确和错误句子的注释与数据集中黄金标准的性别注释进行了比较，四位翻译员的 Cohen's κ 分数在 93.17 到 93.27 之间，确认除了某些语言的语言特性之外（例如，西班牙语中的 "victim" 总是阴性名词，因此无论上下文中的代词是什么都会采用同一形式），注释者都一致同意哪些句子应该以哪种性别正确翻译。

E 整体模型性能

表 7 以 COMET 指标 [5] (Rei et al., 2022) 展示了在包含目标语言 (Callison-Burch et al., 2012; Bojar et al., 2013, 2014; Koehn et al., 2023; Haddow et al., 2024) 的 WMT 数据集上使用的模型性能。这些模型在一块 NVIDIA TU102 GPU 上运行。

[5]https://huggingface.co/Unbabel/wmt22-comet-da

Model	OPUS-MT				deb-OPUS-MT				M2M100			
Dataset	es	fr	uk	ru	es	fr	uk	ru	es	fr	uk	ru
newstest2012	84.52	82.21	–	–	84.47	82.22	–	–	71.25	71.97	–	–
newstest2013	85.28	83.45	–	–	85.24	83.44	–	–	72.84	72.57	–	–
newstest2014	–	85.01	–	87.44	–	85.00	–	87.44	–	74.63	–	72.29
wmttest2023	–	–	74.58	79.02	–	–	74.58	79.02	–	–	56.49	55.93
wmttest2024	–	–	66.99	71.64	–	–	67.00	71.63	–	–	49.20	47.63
mean	84.90	83.56	70.79	79.37	84.86	83.55	70.79	79.36	72.05	73.06	52.85	58.62

Table 7: 使用模型在 WMT 测试集上的 COMET 分数。

F ANOVA 结果

表格 8 展示了 S3E 、SE 和 GE 指标的 ANOVA 结果。表格 9 展示了未对 $\mathcal{H}$ 值进行归一化的 ANOVA 结果。

Lang.	Model	Names	Recency		Implicit Causality				Stereotype				Subject		Context				Default M		Ambiguity
			F	M	SF	SM	OF	OM	SF	SM	OF	OM	F	M	SF	SM	OF	OM	S	O	
S3E																					
	OPUS-MT	0.41	0.41	-0.05	0.25	-0.19	0.24	-0.33	0.06	0.10	0.13	0.17	0.38	-0.21	0.24	-0.31	0.29	-0.13	N/A	N/A	-0.18
es	deb-OPUS-MT	-0.05	0.30	-0.11	0.27	-0.20	0.16	-0.38	0.05	0.14	0.08	0.05	0.49	-0.14	0.39	-0.24	0.14	-0.21	N/A	N/A	-0.10
	M2M100	0.14	0.33	-0.11	0.29	-0.42	0.28	-0.25	0.18	0.27	0.12	0.07	0.51	-0.04	0.52	-0.08	0.04	-0.35	N/A	N/A	-0.11
	OPUS-MT	0.54	0.42	0.16	0.05	-0.34	0.06	-0.24	0.43	0.45	0.26	0.21	0.16	-0.14	0.16	-0.17	0.20	-0.04	N/A	N/A	-0.29
fr	deb-OPUS-MT	0.19	0.22	0.02	0.05	-0.25	0.17	-0.11	0.23	0.32	0.23	0.13	0.24	-0.03	0.31	-0.08	0.01	-0.14	N/A	N/A	-0.12
	M2M100	-0.12	0.11	-0.23	0.50	0.09	0.49	0.03	-0.03	0.11	0.21	0.08	0.70	0.31	0.47	0.05	-0.08	-0.40	N/A	N/A	0.06
	OPUS-MT	-0.27	0.00	-0.11	0.26	-0.02	0.19	0.13	-0.11	0.17	0.03	-0.14	0.27	0.21	0.22	0.13	-0.11	-0.23	N/A	N/A	0.06
uk	deb-OPUS-MT	-0.43	-0.06	-0.12	0.41	0.00	0.18	0.07	-0.24	0.01	0.01	-0.17	0.23	0.17	0.16	0.10	-0.11	-0.17	N/A	N/A	0.09
	M2M100	0.08	-0.04	-0.18	0.27	0.02	0.33	0.18	0.19	0.37	-0.03	-0.17	0.53	0.34	0.50	0.26	-0.30	-0.42	N/A	N/A	0.11
	OPUS-MT	0.01	-0.28	-0.41	0.00	-0.12	0.08	-0.10	-0.19	-0.20	-0.41	-0.39	0.07	-0.03	0.14	0.03	-0.10	-0.24	-0.40	0.04	0.35
ru	deb-OPUS-MT	0.23	-0.10	-0.17	0.05	0.03	0.05	-0.03	-0.10	-0.04	-0.11	-0.14	0.09	0.04	-0.13	0.03	-0.08	-0.13	-0.26	0.00	0.13
	M2M100	-0.74	-0.16	-0.21	0.10	-0.04	-0.12	-0.25	0.12	0.12	-0.20	-0.18	0.06	-0.02	0.08	0.01	-0.07	-0.11	-0.20	-0.22	0.18
SE																					
	OPUS-MT	-1.64	0.19	-0.08	0.13	-0.19	0.09	-0.29	-0.01	0.02	0.06	0.07	0.28	-0.12	0.17	-0.22	0.15	-0.08	N/A	N/A	-0.05
es	deb-OPUS-MT	-0.06	0.27	0.13	-0.01	-0.07	0.00	-0.19	0.06	0.10	0.16	0.12	-0.01	-0.18	-0.01	-0.19	0.20	0.08	N/A	N/A	-0.20
	M2M100	-1.71	0.23	-0.04	0.20	-0.29	0.17	-0.16	0.10	0.18	0.15	0.08	0.28	-0.06	0.25	-0.09	0.07	-0.16	N/A	N/A	-0.10
	OPUS-MT	-1.59	0.19	0.00	0.09	-0.25	0.05	-0.17	0.23	0.25	0.15	0.11	0.11	-0.12	0.10	-0.15	0.14	-0.04	N/A	N/A	-0.09
fr	deb-OPUS-MT	-0.01	0.24	0.08	0.11	-0.17	0.04	-0.10	0.23	0.20	0.19	0.18	-0.06	-0.22	0.01	-0.15	0.18	0.03	N/A	N/A	-0.16
	M2M100	-0.15	0.29	0.16	0.03	-0.05	0.02	-0.14	0.27	0.24	0.27	0.30	0.01	-0.14	0.01	-0.15	0.19	0.06	N/A	N/A	-0.22
	OPUS-MT	-1.54	0.00	-0.12	0.21	-0.05	0.10	0.02	-0.09	0.05	0.01	-0.08	0.13	0.04	0.10	0.02	0.00	-0.14	N/A	N/A	0.06
uk	deb-OPUS-MT	0.09	0.12	-0.06	0.13	-0.07	0.08	0.01	-0.02	0.08	0.15	0.09	0.06	-0.11	-0.07	-0.10	0.13	0.06	N/A	N/A	0.09
	M2M100	-1.72	-0.04	-0.14	0.20	0.03	0.15	0.08	0.05	0.16	-0.02	-0.11	0.22	0.10	0.18	0.09	-0.09	-0.19	N/A	N/A	0.09
	OPUS-MT	-1.50	-0.17	-0.25	0.21	0.07	0.00	-0.02	-0.13	-0.18	-0.25	-0.28	0.03	-0.03	0.10	0.00	-0.06	-0.14	-0.32	-0.02	0.21
ru	deb-OPUS-MT	-0.06	0.04	0.01	0.11	0.05	0.00	-0.06	-0.22	-0.13	0.07	-0.01	-0.11	-0.15	-0.05	-0.11	0.09	0.07	-0.24	-0.12	0.03
	M2M100	-1.57	-0.09	-0.14	0.11	0.10	0.04	-0.18	-0.01	0.00	-0.10	-0.10	0.00	-0.06	0.03	-0.01	-0.01	-0.07	-0.12	-0.18	0.11
GE																					
	OPUS-MT	0.02	0.30	0.04	-0.17	0.16	-0.10	-0.14	0.11	0.15	0.17	0.12	0.19	-0.12	0.19	-0.20	0.15	-0.06	N/A	N/A	0.17
es	deb-OPUS-MT	-0.04	0.27	0.06	0.16	-0.10	0.15	-0.12	0.09	0.17	0.17	0.09	0.21	-0.09	0.24	-0.16	0.09	-0.06	N/A	N/A	-0.16
	M2M100	-0.10	0.16	0.00	0.08	-0.10	0.09	-0.08	0.05	0.07	0.10	0.09	0.09	-0.08	0.08	-0.16	0.12	-0.01	N/A	N/A	-0.08
	OPUS-MT	0.01	0.20	0.05	0.12	-0.06	0.07	-0.11	0.09	0.13	0.12	0.08	0.09	-0.10	0.06	-0.19	0.15	0.03	N/A	N/A	-0.13
fr	deb-OPUS-MT	-0.05	0.18	0.04	-0.11	0.08	0.08	-0.11	0.05	0.12	0.12	0.06	0.11	-0.09	0.10	-0.17	0.11	0.01	N/A	N/A	-0.11
	M2M100	-0.02	0.19	0.02	0.12	-0.10	0.09	-0.12	0.08	0.10	0.11	0.07	0.08	-0.11	0.04	-0.19	0.17	0.02	N/A	N/A	-0.11
	OPUS-MT	-0.04	0.05	-0.04	0.07	-0.03	0.03	-0.07	0.06	0.06	0.02	0.02	0.02	-0.06	-0.11	-0.15	0.16	0.06	N/A	N/A	0.01
uk	deb-OPUS-MT	-0.03	0.05	0.02	0.07	-0.03	0.00	-0.06	0.06	0.07	0.03	-0.03	-0.02	-0.05	-0.16	-0.17	0.17	0.14	N/A	N/A	0.03
	M2M100	-0.04	0.06	-0.04	0.09	-0.05	0.05	-0.06	0.07	0.08	0.01	0.01	0.03	-0.07	-0.07	-0.18	0.16	0.06	N/A	N/A	0.01
	OPUS-MT	-0.01	0.00	-0.03	-0.02	-0.01	0.00	-0.04	0.03	0.03	-0.04	-0.06	-0.01	-0.04	-0.11	-0.13	0.12	0.08	-0.04	0.00	-0.02
ru	deb-OPUS-MT	-0.02	-0.01	-0.02	-0.05	-0.05	-0.01	-0.02	0.01	0.02	-0.04	-0.05	-0.03	-0.04	-0.15	-0.15	0.13	0.12	-0.05	-0.01	-0.02
	M2M100	0.02	0.07	0.02	-0.01	-0.06	-0.01	-0.07	0.09	0.07	0.02	0.04	-0.01	-0.06	-0.05	-0.11	0.10	0.06	-0.04	-0.03	0.04

Table 8: ANOVA 结果：偏见线索（女性化、男性化、主体和客体）对 norm- $\mathcal{H}$ （S3E）、norm- $\mathcal{H}$ （SE）和 norm- $\mathcal{H}$ （GE）的单一效应。数值对应于效应系数（与参考组的偏差）。粗体表示统计显著性（$p < 0.05$）。数值的符号表示该变量的存在是否增加（正值）或减少（负值）包含该变量值的组的平均 $\mathcal{H}$ 。参考组为 N，所有列均为该组，除以下情况：对于“名字”，参考组为“无名字”；对于“默认 M”，参考组为“无默认”；对于“模糊性”，参考组为“无歧义”。

Lang.	Model	Names	Recency		Implicit Causality				Stereotype				Subject		Context				Default M		Ambiguity
			F	M	SF	SM	OF	OM	SF	SM	OF	OM	F	M	SF	SM	OF	OM	S	O	
S3E																					
	OPUS-MT	12.12	47.0	83.53	-2.45	20.67	-2.44	27.76	38.98	49.12	44.16	37.6	0.23	31.36	7.95	42.89	-22.44	14.64	N/A	N/A	-65.29
es	deb-OPUS-MT	-2.35	36.52	79.64	-58.1	-7.06	18.24	-8.44	31.09	33.94	42.22	36.85	32.48	-1.86	34.19	6.06	48.72	-29.24	N/A	N/A	14.06
	M2M100	0.19	0.54	0.02	-0.28	0.38	-0.27	0.37	-0.26	0.18	0.25	0.25	0.17	0.31	-0.25	0.31	-0.29	0.31	N/A	N/A	-0.17
	OPUS-MT	0.12	0.41	0.06	0.32	-0.15	0.23	-0.17	0.18	0.25	0.23	0.14	0.26	-0.16	0.32	-0.19	0.16	-0.14	N/A	N/A	-0.24
fr	deb-OPUS-MT	-0.02	0.26	0.06	-0.16	0.23	-0.07	0.15	-0.1	0.12	0.21	0.18	0.08	0.21	-0.08	0.25	-0.12	0.05	N/A	N/A	-0.08
	M2M100	-0.12	0.11	-0.23	0.06	0.5	0.09	0.49	0.03	-0.03	0.11	0.21	0.08	0.7	0.31	0.47	0.05	-0.08	N/A	N/A	-0.4
	OPUS-MT	0.12	0.22	0.03	0.16	-0.07	0.09	-0.11	0.15	0.18	0.14	0.1	0.09	-0.1	0.05	-0.1	0.16	-0.04	-0.09	-0.05	-0.12
uk	deb-OPUS-MT	0.83	-10.2	24.93	-7.38	-19.32	18.37	-14.82	27.31	5.26	9.67	8.87	5.29	-15.13	23.85	-10.15	44.14	-26.98	1.7	-2.21	-0.34
	M2M100	0.14	0.3	0.06	-0.18	0.2	-0.16	0.12	-0.13	0.19	0.19	0.16	0.15	0.1	-0.15	0.05	-0.17	0.22	-0.02	-0.1	-0.04
	OPUS-MT	15.38	4.07	52.95	-27.98	17.38	-24.19	20.45	11.79	12.51	12.13	13.6	-24.94	22.11	-28.52	41.31	-18.56	23.26	-0.79	-1.99	-28.53
ru	deb-OPUS-MT	0.23	-0.1	-0.17	0.13	0.05	0.03	0.05	-0.03	-0.1	-0.04	-0.11	-0.14	0.09	0.04	0.13	0.03	-0.08	-0.13	-0.26	-0.0
	M2M100	0.08	-0.07	-0.05	0.06	-0.02	-0.01	-0.02	-0.0	-0.06	-0.07	-0.05	-0.03	-0.01	0.0	-0.02	0.0	-0.01	-0.0	0.01	0.03
SE																					
	OPUS-MT	12.12	47.0	83.53	-2.45	20.67	-2.44	27.76	38.98	49.12	44.16	37.6	0.23	31.36	7.95	42.89	-22.44	14.64	N/A	N/A	-65.29
es	deb-OPUS-MT	-0.08	-0.06	-0.05	0.05	-0.01	0.0	-0.01	-0.0	-0.04	-0.05	-0.04	-0.03	-0.01	0.0	-0.01	0.01	-0.02	N/A	N/A	-0.01
	M2M100	0.08	-0.07	-0.06	0.07	-0.01	-0.0	-0.02	-0.01	-0.05	-0.05	-0.04	-0.04	-0.01	0.0	-0.02	0.01	-0.02	N/A	N/A	-0.01
	OPUS-MT	16.01	47.22	80.76	-0.14	19.01	-8.69	26.55	31.41	35.76	33.4	31.98	-1.93	29.87	2.67	35.97	-15.81	17.83	N/A	N/A	-64.01
fr	deb-OPUS-MT	-0.07	-0.05	-0.04	0.04	-0.0	0.02	-0.01	0.0	-0.03	-0.04	-0.03	-0.03	-0.01	0.01	-0.01	0.01	-0.02	N/A	N/A	-0.01
	M2M100	0.14	0.38	0.04	-0.21	0.35	-0.15	0.23	-0.17	0.15	0.2	0.19	0.13	0.23	-0.16	0.24	-0.19	0.19	N/A	N/A	-0.12
	OPUS-MT	-0.01	-0.01	-0.01	-0.03	-0.02	0.0	0.0	-0.03	-0.03	-0.01	-0.01	-0.0	0.0	-0.01	0.01	-0.0	-0.0	0.02	0.03	0.01
uk	deb-OPUS-MT	0.0	0.09	0.02	-0.05	0.01	-0.06	0.06	-0.04	0.12	0.16	0.08	0.04	0.04	-0.02	0.08	0.02	0.01	-0.06	-0.1	-0.03
	M2M100	0.14	0.3	0.06	-0.18	0.2	-0.16	0.12	-0.13	0.19	0.19	0.16	0.15	0.1	-0.15	0.05	-0.17	0.22	-0.02	-0.1	-0.04
	OPUS-MT	15.38	4.07	52.95	-27.98	17.38	-24.19	20.45	11.79	12.51	12.13	13.6	-24.94	22.11	-28.52	41.31	-18.56	23.26	-0.79	-1.99	-28.53
ru	deb-OPUS-MT	1.34	-3.62	46.99	-21.71	-33.31	16.73	-27.33	19.72	9.4	7.79	6.6	9.66	-28.51	22.12	-24.71	50.74	-26.37	15.86	-0.28	-0.68
	M2M100	20.01	18.76	50.38	-34.59	-11.85	5.91	-16.63	13.63	14.94	15.4	14.21	15.2	-17.46	14.68	-15.4	30.42	-10.35	16.48	-0.74	-3.23
GE																					
	OPUS-MT	-0.0	-0.02	-0.01	-0.01	0.01	-0.01	0.0	-0.02	-0.02	-0.01	-0.01	-0.01	0.01	-0.0	0.01	-0.01	-0.0	N/A	N/A	0.02
es	deb-OPUS-MT	-2.35	36.52	79.64	-58.1	-7.06	18.24	-8.44	31.09	33.94	42.22	36.85	32.48	-1.86	34.19	6.06	48.72	-29.24	N/A	N/A	14.06
	M2M100	0.19	0.54	0.02	-0.28	0.38	-0.27	0.37	-0.26	0.18	0.25	0.25	0.17	0.31	-0.25	0.31	-0.29	0.31	N/A	N/A	-0.17
	OPUS-MT	0.06	-0.05	-0.04	0.01	0.02	-0.01	0.01	-0.04	-0.04	-0.03	-0.02	-0.0	0.01	-0.01	0.01	-0.02	-0.01	N/A	N/A	0.04
fr	deb-OPUS-MT	-0.05	0.18	0.04	-0.11	0.08	-0.11	0.08	-0.11	0.05	0.12	0.12	0.06	0.11	-0.09	0.1	-0.17	0.11	N/A	N/A	0.01
	M2M100	0.14	0.38	0.04	-0.21	0.35	-0.15	0.23	-0.17	0.15	0.2	0.19	0.13	0.23	-0.16	0.24	-0.19	0.19	N/A	N/A	-0.12
	OPUS-MT	0.12	0.22	0.03	0.16	-0.07	0.09	-0.11	0.15	0.18	0.14	0.1	0.09	-0.1	0.05	-0.1	0.16	-0.04	-0.09	-0.05	-0.12
uk	deb-OPUS-MT	0.0	0.09	0.02	-0.05	0.01	-0.06	0.06	-0.04	0.12	0.16	0.08	0.04	0.04	-0.02	0.08	0.02	0.01	-0.06	-0.1	-0.03
	M2M100	0.14	0.3	0.06	-0.18	0.2	-0.16	0.12	-0.13	0.19	0.19	0.16	0.15	0.1	-0.15	0.05	-0.17	0.22	-0.02	-0.1	-0.04
	OPUS-MT	0.09	0.17	0.05	0.01	-0.11	0.05	-0.08	0.08	0.13	0.11	0.06	0.04	-0.08	-0.02	-0.06	0.14	-0.01	-0.08	-0.03	-0.11
ru	deb-OPUS-MT	-0.02	-0.01	-0.02	0.02	-0.05	-0.05	-0.01	-0.02	0.01	0.02	-0.04	-0.05	-0.03	-0.04	-0.15	-0.15	0.13	0.12	-0.05	-0.01
	M2M100	0.15	0.29	0.08	-0.18	0.13	-0.1	0.09	-0.12	0.16	0.21	0.15	0.11	0.07	-0.13	0.08	-0.05	0.15	-0.08	-0.1	-0.02

Table 9: ANOVA 结果（无归一化）：偏倚线索（女性、男性、主语和宾语）对 $\mathcal{H}$ （S3E）、$\mathcal{H}$ （SE）和 $\mathcal{H}$ （GE）的单一效应。数值对应效应系数（与参考组的偏差）。黑体字表示统计显著性（$p < 0.05$）。数值的符号表示变量的存在是增加（正数）还是减少（负数）包含给定变量值的组的平均 $\mathcal{H}$ 。参考组是 N，除了：Name 为“无名”，Default M 为“无默认值”，Ambiguity 为“明确”。

G 性别准确性

表格 10 在性别准确性方面给出了比表格 2 更细化的结果，即按数据集的子集细分结果。模糊列中的结果没有意义上的可解释性，因为单一的性别标注无法捕获模型的真实理想行为，尤其是当模糊情况的金标大多是“中性”时，而中性并不常用作该研究中所用语言的有生命物体的语法性别。在俄语的情况下，模糊子集上性能的提高实际上反映了模型选择了阳性形式，而这些阳性形式由于阳性形式通常在两性中作为默认选择而被形态分析器标记为“中性”，正如在第 4 节中讨论的。

Lang.	Model	All	Pro	Anti	Unamb.	Amb.
es	OPUS-MT	55.20	67.95	52.10	67.95	33.96
	deb-OPUS-MT	55.69	68.13	52.95	68.13	34.39
	M2M100	55.68	70.77	51.17	70.77	32.40
fr	OPUS-MT	52.05	64.27	46.55	64.27	37.25
	deb-OPUS-MT	52.98	64.79	48.10	64.79	37.75
	M2M100	50.95	61.66	47.57	61.66	34.84
uk	OPUS-MT	38.65	45.34	34.20	45.34	33.75
	deb-OPUS-MT	38.95	46.12	34.15	46.12	33.74
	M2M100	40.97	47.76	36.81	47.76	35.20
ru	OPUS-MT	39.50	48.57	33.27	48.57	33.24
	deb-OPUS-MT	39.50	48.42	33.38	48.42	33.33
	M2M100	41.01	48.49	36.81	48.49	33.81

Table 10: 在模型中，关于 WINOMT 的亲/反刻板印象和模糊案例中的性别准确性比较。

H 用人工翻译进行质量评估

表格 11 展示了本研究使用的模型在 100 个人工标注实例上的结果。对于无歧义的情况，我们使用单一参考，而对于有歧义的情况，我们通过取两个可接受译文的最大 COMET 分数来计算表现。

Lang.	Model	All	Pro	Anti	Unamb.	Amb.
es	OPUS-MT	81.35	85.80	83.37	84.55	75.14
	deb-OPUS-MT	81.31	85.62	83.43	84.49	75.14
	M2M100	79.56	84.44	81.29	82.82	73.25
fr	OPUS-MT	77.63	82.23	81.16	81.66	70.47
	deb-OPUS-MT	77.69	82.41	80.88	81.60	70.73
	M2M100	76.24	80.69	78.65	79.61	70.25
uk	OPUS-MT	80.56	85.57	82.73	84.10	73.69
	deb-OPUS-MT	80.19	84.85	82.13	83.45	73.86
	M2M100	81.27	85.89	84.09	84.96	74.10
ru	OPUS-MT	82.53	86.20	84.99	85.58	76.86
	deb-OPUS-MT	82.76	86.46	85.27	85.85	77.02
	M2M100	81.71	86.39	84.06	85.21	75.21

Table 11: 在被手动翻译的 100 个句子中，各个模型在整体、正/反刻板印象和模糊案例中 COMET 得分的比较。

Language	Model	LogProb (Correct)	LogProb (Incorrect)	S3E I (Correct)	S3E I (Incorrect)	SE I (Correct)	SE I (Incorrect)	GE I (Correct)	GE I (Incorrect)
ES	OPUS-MT	-149.7	-149.78	7.83	8.88	0.3	0.35	0.33	0.3
	deb-OPUS-MT	-149.19	-149.01	8.08	9.16	0.29	0.31	0.35	0.33
	M2M100	-226.61	-227.29	23.61	26.03	0.41	0.4	0.42	0.43
FR	OPUS-MT	-197.1	-195.11	9.18	9.89	0.73	0.72	0.24	0.29
	deb-OPUS-MT	-196.98	-195.09	9.18	9.85	0.48	0.52	0.26	0.33
	M2M100	-283.91	-281.71	186.42	194.52	0.49	0.4	0.43	0.48
UK	OPUS-MT	-161.98	-161.14	147.6	152.15	0.6	0.54	0.22	0.22
	OPUS-MT-debiased	-161.46	-160.68	150.52	153.76	0.49	0.47	0.23	0.23
	M2M100	-241.0	-241.49	204.72	211.9	0.28	0.24	0.23	0.25
RU	OPUS-MT	-170.72	-170.9	32.14	33.15	0.38	0.43	0.08	0.19
	deb-OPUS-MT	-170.58	-170.75	32.31	33.27	0.32	0.37	0.06	0.17
	M2M100	-220.25	-220.78	218.11	219.06	0.45	0.4	0.16	0.3

Table 12: 各模型和语言中的对数概率和意外性测量

COMET Unambiguous	COMET All	Gender Acc	Delta S	Delta H
deb-OPUS-MT-RU	OPUS-MT-ES	M2M100-ES	M2M100-ES	M2M100-RU
OPUS-MT-RU	deb-OPUS-MT-ES	deb-OPUS-MT-ES	deb-OPUS-MT-ES	M2M100-UK
M2M100-RU	OPUS-MT-FR	OPUS-MT-ES	OPUS-MT-ES	OPUS-MT-UK
M2M100-UK	deb-OPUS-MT-FR	deb-OPUS-MT-FR	deb-OPUS-MT-FR	deb-OPUS-MT-UK
OPUS-MT-ES	OPUS-MT-RU	OPUS-MT-FR	M2M100-FR	OPUS-MT-RU
deb-OPUS-MT-ES	deb-OPUS-MT-RU	M2M100-FR	OPUS-MT-FR	deb-OPUS-MT-RU
OPUS-MT-UK	M2M100-FR	OPUS-MT-RU	deb-OPUS-MT-RU	deb-OPUS-MT-ES
deb-OPUS-MT-UK	M2M100-ES	M2M100-RU	M2M100-RU	OPUS-MT-ES
OPUS-MT-ES	OPUS-MT-UK	deb-OPUS-MT-RU	deb-OPUS-MT-UK	deb-OPUS-MT-FR
OPUS-MT-FR	deb-OPUS-MT-UK	M2M100-UK	OPUS-MT-UK	M2M100-FR
deb-OPUS-MT-FR	M2M100-RU	deb-OPUS-MT-UK	M2M100-UK	M2M100-ES
M2M100-FR	M2M100-UK	OPUS-MT-UK	OPUS-MT-RU	OPUS-MT-FR

Table 13: 模型在五个评价指标上的排名。

I 对数概率和意外度分数

表 12 显示了不含糊实例在 WINOMT 中正确和错误翻译的对数概率和意外度分数，以及它们之间的相对差异。

J 不同指标的排名

表 13 列出了根据本研究中使用的各种指标对模型的排名。

K 秩相关

表格 14 展示了一方面 ΔI 和 Δ 对数概率之间的相关性得分，另一方面展示了性别准确性得分。

Correlation	Metric	Statistic	p-value
Spearman	ΔI (S3E)	-0.78	0.00
	ΔI (SE)	-0.37	0.24
	ΔI (GE)	0.27	0.00
	ΔLog prob	0.11	0.73
Kendall	ΔI (S3E)	-0.58	0.01
	ΔI (SE)	-0.27	0.25
	ΔI (GE)	0.23	0.00
	ΔLog prob	0.09	0.74

Table 14: 不同不确定性度量下的 ΔI 与对数概率之间，以及与性别准确率之间的 Spearman 和 Kendall 相关性。统计上显著的相关性（$p < 0.05$）以粗体显示。

Language	Model	S3E			SE			GE		
		Unamb.	Amb.	$\Delta\mathcal{H}$	Unamb.	Amb.	$\Delta\mathcal{H}$	Unamb.	Amb.	$\Delta\mathcal{H}$
ES	OPUS-MT	1.23	1.12	0.09	0.33	0.22	0.33	0.21	0.13	0.38
	deb-OPUS-MT	0.97	0.89	0.08	0.41	0.25	0.39	0.24	0.16	0.33
	M2M100	1.79	1.45	0.19	0.46	0.17	0.63	0.25	0.09	0.64
FR	OPUS-MT	1.79	1.43	0.20	0.57	0.40	0.30	0.23	0.08	0.65
	deb-OPUS-MT	1.21	1.08	0.11	0.64	0.50	0.22	0.23	0.09	0.61
	M2M100	3.22	2.78	0.14	0.56	0.29	0.48	0.25	0.17	0.32
UK	OPUS-MT	1.96	2.16	-0.10	0.40	0.39	0.03	0.20	0.14	0.30
	deb-OPUS-MT	1.98	2.15	-0.09	0.44	0.41	0.07	0.22	0.15	0.32
	M2M100	2.05	2.28	-0.11	0.37	0.41	-0.11	0.18	0.16	0.11
RU	OPUS-MT	1.56	1.68	-0.08	0.38	0.32	0.16	0.12	0.03	0.75
	deb-OPUS-MT	1.05	0.97	0.08	0.43	0.42	0.02	0.12	0.06	0.50
	M2M100	1.83	2.29	-0.25	0.50	0.29	0.42	0.15	0.10	0.33

Table 15: 不同模型和语言的 $\mathcal{H}$ 分数，以及在非模糊和模糊条件下的相对差异（$\Delta\mathcal{H}$）。

L 熵分数

表 15 展示了在本研究中使用的不同不确定性量化指标下，不明确和有歧义设置之间的 $\mathcal{H}$ 分数及其相对差异。