Student*t*分布：观测值稀少时如何测量评估者间信度

Serge Gladkoff [1*]，**Lifeng Han**[2*]，和**Goran Nenadic**[21] Logrus Global，翻译与本地化[2] 曼彻斯特大学，
英国 lifeng.han, g.nenadic @ manchester.ac.uk
serge.gladkoff @ logrusglobal.com
* 共同第一作者

arXiv:2303.04526v2 [cs.CL] 9 Jul 2023

摘要

在自然语言处理（NLP）领域，我们始终依赖人工判断，将其视为黄金标准质量评估方法。然而，如何更好地评估某些评估任务（例如翻译质量评估（TQE））中的评估者间信度（IRR）水平，一直存在争论，尤其是在数据样本（观测值）非常稀少时。实际上，从业者需要能够根据一位、两位或最多三位人类语言学家的判断，评估人工机器翻译质量评估的信度。在这项工作中，我们首先介绍一种鲜为人知的方法，用于在仅有一个数据（评估）点时估算测量值的置信区间。接着，我们以两个由人工生成的观测分数为例，介绍“Student 的*t*分布”，并说明如何仅利用这两个数据点测量 IRR 分数，以及计算质量评估的置信区间（CI）。我们定量分析了增加观测值如何显著提高评估置信度，即使只额外增加一个观测值也能如此。我们鼓励从业者和研究人员在所有评估中报告 IRR 分数及其置信区间，例如尽可能采用 Student 的*t*分布方法；从而使自然语言处理评估更有意义、更透明、更可信。

1 引言

人工评估一直是衡量自然语言处理（NLP）系统输出质量的黄金标准（Han et al., 2021; Freitag et al., 2021; Gladkoff and Han, 2022）。这适用于许多子任务，包括机器翻译（MT）（Han et al., 2020; Han, 2022a; Charalampidou and Gladkoff, 2022; MI- LAD, 2022）、文本摘要（Bhandari et al., 2020; Latif et al., 2009）、问答（Al-rdahi et al., 2020）、信息抽取（Wu et al., 2022; Nenadic et al., 2004）和预测（Yang et al., 2009），以及社交媒体、生物医学和临床领域的知识表示等领域应用（Miloˇsevi´c et al., 2019; Yang et al., 2021; Krauthammer and Nenadic, 2004）。尽管如此，人工评估一直受到批评，并引发了关于其信度的争论，尤其是在缺乏严格界定的流程时（Han, 2022b; Han and Gladkoff, 2022; Graham et al., 2017）。即使纳入了质量控制和明确的指南等因素，人工评估结果在不同个体之间仍可能存在很大差异，因为判断具有主观性，并会受到背景、个性、文化等因素的影响。

自然地，人类评估的置信水平成为此类工作有效性的关键。有一些研究致力于从统计学角度测量人类评估的置信水平，例如 Gladkoff 等人（2022）近期采用蒙特卡洛抽样模拟开展的研究。然而，这类统计测量仍需要足量的数据点或观测值作为依据。当实验获得的观测值数量有限时，如何恰当地测量置信水平仍是一个具有挑战性的问题。解决这一问题的方法之一，是计算评估者间一致性水平和评估者间信度（IRR）得分。历史上已有一些评估者间可靠性测量指标，包括 Cohen κ系数（Cohen，1960）和 Krippendorff α系数（Krippendorff, 1987、2011）。如何—

ever，正如统计抽样所面临的最后一个问题一样，Cohen κ系数和 Krippendorff α系数都需要一定数量的样本数据来进行概率计算；当观测值确实稀缺时，例如只有一个、两个或少数几个数据点，这就会带来麻烦。此外，针对 Cohen κ系数不理想的预测结果，现有研究也提出了批评。例如，Delgado 和 Tibau（2019）举例说明了 Kappa 如何为表现更差的分类器产生更高的得分。

在本研究中，我们考察人工评估的观测值极其有限（例如，仅有一两个数值）的情形，并探讨可能的解决方案。这项研究源于翻译与本地化行业的实际需求。从业者往往需要根据一位语言学家的判断，或至多两到三位语言学家的评估，来确定人工机器翻译（MT）质量评估的信度。

我们从一次观测入手，介绍一种置信度估计方法。该方法借鉴自Abbott和Rosenblatt（1962）以及Furnival等人（1989），并由Valentine等人（1991）应用于森林研究。随后，我们讨论基于一次观测进行置信度估计存在的问题，以及这种方法为何不够可靠。接着，我们举例说明一项基于人工评估获得两次观测值的机器翻译评估。在这个案例研究中，我们介绍如何应用**Student的*t*分布**来衡量评估者间信度，并详细解读公式、提供操作指导。我们还进一步说明如何运用该方法计算置信区间（CIs）。我们讨论了使用两个观察数据点和Student*t*分布所带来的显著改进，以及由此缩小的置信区间。最后，我们建议研究人员也将Student*t*分布应用于其他自然语言处理任务，乃至自然语言处理任务之外的领域。

本文其余部分安排如下：第 2 节综述了不同领域中与我们关于测量评估者间信度及置信区间的研究相关的工作；第 3 节介绍一个仅有单个观测值的案例研究；第 4 节进一步讨论两个及更多观测值的情形，并引入 Student 的*t*分布；第 5 节总结全文，并讨论未来工作。

2 相关工作

关于一致性测量，Cohen 的 Kappa 指标由 Cohen（1960）定义如下：

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

其中，p_o 用于表示评审者之间达成一致时的比例（单位占比），而 p_e 表示随机情况下预期的一致性。根据频数表达式，Kappa 值可按以下方式计算：

$$\kappa = \frac{f_o - f_e}{N - f_e} \tag{2}$$

换言之，Kappa 值反映的是扣除随机一致性后的整体一致程度（或一致比例）。在理想情况下，所有评审者彼此一致，即 p_o 值等于 1 时，Kappa 值将为 1 （ Cantor，1996 年）。然而，如果评审者完全不一致，即 p_o 的值几乎与随机一致性的值 p_e相同，那么 Kappa 值将接近或等于 0。不过，当一致程度甚至低于随机一致性时，Kappa 值也可能为负数，例如在上述方程中，$p_o - p_e$ 或 $f_o - f_e$ 的值为负数。如前文所述，要正确估算指标分数，Kappa 值需要一定数量的观测值。

在考察众包人工评估领域的评估者间可靠性测量时，Wong 等人（2021）指出，传统的 Krippendorff α系数或 Cohen κ系数阈值（例如高于 0.6）并不理想，因为这些阈值忽视了众包工作者的文化差异和个体差异。他们提出了一种基于 Cohen κ系数的跨重复实验可靠性方法，并利用多达 400 万个数据点，对面部表情的人工判断进行了方法测试。

在自然语言处理和机器翻译领域，Alekseeva 等人（2021）和 Gladkoff 等人（2022）采用蒙特卡洛模拟分析方法生成更多样本，以便在用于人工评估的样本数量较少时，对判断的置信区间进行统计估计。他们的实验结果表明，至少

需要 100 个、理想情况下 200 个片段，才能使测试集对机器翻译系统输出进行无偏且具有统计显著性的质量评估。

在自然语言处理领域之外，也有人尝试解决测量信度和置信区间时遇到的类似问题。例如，Hallgren（2012）针对心理学领域中涉及多名编码者的情况，结合案例研究，就如何测量评估者间信度进行了教程讲解，采用的是常用的 Cohen κ系数和组内相关系数（ICC）。同样，在教育和心理学领域，Walker 和 G¨oc¸er S¸ahin（2020）开展了一项研究，将项目差异功能（DIF）分析用于测量评估者间信度，并与组间相关系数及 Cohen κ系数进行了比较。

在动物行为研究中，Harvey（2021）提出了评估者间和评估者内信度方面的问题，并讨论了 Cohen κ系数和 Krippendorff α系数的取值。在社会学领域，Belur 等人（2021）报告了一项系统性调查，考察犯罪研究中多名编码者的评估者间信度值报告情况。他们进一步讨论了人为因素如何影响决策，以及报告筛查／编码的准确度、精确度和信度的重要性。

目前已有一些商业软件将评估者间信度集成到统计工具中，例如 SPSS；这些工具已应用于不同领域，包括医学培训评估（Beck 等，2016）。IBM SPSS 使用组内相关系数（ICC）来测量不同评估者群体之间的评估者间信度值 [1]。

然而，据我们所知，目前尚无研究将 Student 的*t*-分布用于自然语言处理应用中的评估者间信度测量，尤其是在翻译质量评估（TQE）领域。

3 关于单次判断

在观察数据极其稀少的情况下，半个多世纪前，Abbott 和 Rosenblatt（1962）从数学角度证明了可以基于单个数据点测量置信区间；此后，Furnival等人（1989）进一步阐述了 Abbott 和 Rosenblatt 的公式，并生成了范围更窄的区间。我们取他们姓名的首字母，将其称为**ARF** 区间，作为名称。 [2]

这种方法在统计学上或许显得有悖直觉，但在此值得一提，尤其是生产决策往往基于单次质量测量。这里出现了一个耐人寻味的悖论：许多统计学家会认为，仅凭一次测量不可能确定置信区间，而项目经理却常常依靠单个翻译质量评估（TQE）值来做出决策。实际上，可以对单次测量的可靠性得出结论，但这需要补充信息；例如，在翻译行业中，可以参考已知供应商以往的表现。在 ARF 区间计算方法中，这些额外数据也来自基于实验的先验知识，或基于理论的数值；这些信息虽在测量之外，却源于项目背景。有趣的是，项目经理依据单次测量值做决策时，运用的也是类似的直觉。因此，探究哪些数学原理能够阐明这一情境，颇值得玩味。

置信区间的宽度反映了实验的不确定性，也就是说，区间越宽，对实验设置的了解就越少。

相对较窄的置信区间表明实验条件受控，例如，下列标准化变换公式中的正态分布：

$$\mathrm{z} = \frac{y - \mu}{\sigma} \tag{3}$$

其中，y、μ 和 σ 分别是随机变量、均值和标准差。参数 z 表示标准化变量。

对于只有一次观测的情况，设 $\hat{\mu}$ 为测量之前且独立于测量已知的固定值，y 为实验测量值。Furnival等人（1989年）给出了以下计算区间：

$$ARF = \frac{y + \hat{\mu}}{2} \pm k|y - \hat{\mu}| \tag{4}$$

[1] https://www.ibm.com/uk-en/spss

[2] 在 Rodriguez 于 1996年 进行的另一项关于单次观测置信区间（CI）的研究中，得出了 Herbert Robbins 非参数置信区间，并介绍了另一种用于求取“对数尺度下有限长度的尺度参数”置信区间的方法。

此 ARF区间包含 μ 大于或等于 $1-\alpha$ 的概率，且 α 满足以下与 k 有关的方程：

$$k = \frac{1-\alpha+\sqrt{1-2\alpha}}{2\alpha},\ 0 < \alpha \leq 0.5 \qquad (5)$$

(k, α) 的配对值由 Furnival等人（1989年）在表1中给出。

3.1 使用ARF区间的案例研究

让我们研究一个使用ARF区间进行机器翻译评估的案例。假设某翻译供应商此前已针对若干项目线路接受评估，平均结果为96.3分（$\hat{\mu}$）。随后，另一家供应商给出的翻译质量测量结果较低，为85.2分（y）。单从统计学角度来看，这一测量结果本身有多可靠？它能告诉我们什么？如果假设翻译质量测量值服从正态分布，那么将以往评估结果的平均值作为未来结果的预测值是合理的。下面介绍两种使用ARF区间的方法。

1）如果我们要为真实质量评分构造一个75%的置信区间，就需要根据表1中的说明使用k=1.8，相应的α 值为0.25。使用ARF区间公式可得：

$$ARF = \frac{96.3+85.2}{2} \pm 1.8 \times |96.3-85.2| \quad (6)$$

即为 90.75±19.98。因此，真实质量评分值的 ARF 区间为 [70.77,100]。正如我们从该示例中看到的，75% 的置信区间几乎是测量值本身的一半，即 19.98 的最大偏差占测量结果（90.75）的 22%。尽管单次质量测量的数学精度仅限于这一水平，但明确这些界限仍可能有所帮助。

2）同样，对于真实质量评分的 80% 置信区间（$\alpha = 0.2$），表1中的相应 k 值为 2.31，上述公式给出了以下 ARF 值：

$$ARF = \frac{96.3+85.2}{2} \pm 2.31 \times |96.3-85.2| \qquad (7)$$

即 90.75±25.64。换句话说，质量评分真值的区间为 [65.1, 100]。

由此可见，在 80% 置信区间下，25.64 的最大偏差占测量结果（90.75）的 25%。

从这两个案例研究中可以看出，在翻译质量评估的语境下，更有意思的一点可能是：“置信区间的中点位于先前平均结果与近期较低测量值之间的中间位置。”我们可以总结出一条很好的经验法则：如果单次测量偏离平均值，那么真实值很可能位于平均值与新测量值之间的中点。了解这一点有助于避免对新近产生的异常低单次评分反应过度。

虽然可以测量置信水平，但这个 ARF区间非常宽，更糟的是，它无法通过选择 α 来改善。如表1所示，0.5 是可选区间中最窄的选项。然而，这个数值被认为还不够高，无法产生显著影响。表格右侧显示，α 的值越小，所得区间就越宽。因此，当只有一次观测或一次判断可用时，选择介于 (0.2, 0.25) 之间的 α 值或许是较为折中的做法。从这个案例研究中，我们发现，不建议仅由单次测量构成的评估，因为正如我们的翻译评估示例所示，这样更容易产生偏差。这类测量只能提供粗略的指示性数值，因此必须借助数据科学工具采用数据收集与分析方法，以提高测量本身的质量。这将引出下一节的内容：我们将建议进行第二次质量测量，并讨论在新情况下如何进行测量，以及它会带来多大影响。

4 关于多个观测值

承接上一节，我们**呼吁增加测量点**，用于自然语言处理评估任务，尤其是在语言服务提供商行业中。由于成本方面的考虑，实践中单次观测值仍然十分常见。[3]

例如，当某一方对单次翻译质量测量结果不满意时，可以进行第二次质量测量，以验证第一次测量结果。那么，通过引入额外的观察数据，置信区间能改善多少，

[3]例如，参见语言服务提供商 Logrus Global 的研发报告：https://logrusglobal.com/

	值（α）						
分布	0.50	1/3	0.25	0.2	0.1	0.05	0.01
正态分布（k）	0.05	1.26	1.8	2.31	4.79	9.66	48.39
未知（k）	0.5	1.87	2.91	3.94	8.97	18.99	99

表1：Furnival等人（1989年）给出的正态分布和未知分布中（k, α）的值匹配关系。

呢？要回答这个问题，显而易见的难点是，计算正态分布的数学方差至少需要 20 至 30 个数据点。在样本量小于 30，且未知总体标准差的情况下，可以使用 学生 *t* 分布 ，仅根据 1 至 30 个测量值的数量来评估标准差，例如 2、3 等。

4.1 关于 Student*t* 分布

当样本量（即 观测值）与总体相比非常小时，Student（1908年）设计了 Student 的*t*分布，用于衡量均值误差和估计的置信区间。

当自由度为一个时，Student 的*t*分布的临界值如图1所示，其中包括置信水平、单尾和双尾得分。 [4] 自由度更多时的完整临界值列表见图2。[5] 许多研究者提出了不同的算法，利用手工计算或计算机计算这些临界值，例如 Cheng 和 Fu（1983）的研究，以及 Blair 和 Higgins（1980）的比较研究。

如果使用*T* 作为随机变量，Student 的*t*分布的符号定义如下：

- $T \sim t_{df}$ where $df = n - 1$

其中， df 是自由度， n 是观测值的数量。例如，如果样本量为 $n = 2$，我们计算 $df = 2 - 1 =$1，并将分布写作 $T \sim t_1$。

当标准差未知时，样本均值的误差界定义为：

$$E = (t_{\alpha/2})(\frac{s}{\sqrt{n}}) \tag{8}$$

其中， $t_{\alpha/2}$ 是右侧面积为 $\alpha/2$ 的 t 分数临界值（图2）， s 是观测值（样本）的标准差：

$$s = \sqrt{\frac{\sum(x_i - \bar{\text{x}})^2}{n-1}} \tag{9}$$

其中 $\bar{\text{x}}$是 n 样本的均值：

$$\bar{\text{x}} = (1/n)\sum x_i \tag{10}$$

由此得到的置信区间（CI）是以下范围：

$$CI = (\bar{\text{x}} - E, \bar{\text{x}} + E) \tag{11}$$

4.2 将*t*分布应用于评估者间信度

回顾我们在机器翻译评估中的实验，借助一个语言服务行业实际项目中的案例，我们可以展示如何使用 Student 的*t*分布来衡量评估者间信度值。假设我们在一个翻译评估项目中使用了 Lommel 等人（2014）提出的多维质量指标（MQM） [6] 和专业译员，并得到了两个总体质量分数：满分为 100 分的量表上，QS1=76.85 和 QS2=81.99。[7] 我们可以立即看出，QS2 81.99 比 QS1 76.85 高出 6.7%，反过来，QS1 76.85 比 QS2 81.99 低 6.3%。因此，QS2 与 QS1 的一致性为 93.3%，而 QS1 与 QS2 的一致性为 93.7%。这几乎达到 95% 的一致性，因此在大多数情况下看起来不错。然而，如果通过/不通过阈值为 80，这一差异可能至关重要，影响

[4] https://people.richland.edu/james/lecture/m170/tbl-t.html [5] https://www.stat.purdue.edu/~lfindsen/stat503/t-Dist.pdf

[6] 开源项目 https://themqm.org/ [7] 这是一个来自名为“Whale”的 TQE 工业项目的真实示例。

Conf. Level	50%	80%	90%	95%	98%	99%
One Tail	0.250	0.100	0.050	0.025	0.010	0.005
Two Tail	0.500	0.200	0.100	0.050	0.020	0.010
df = 1	1.000	3.078	6.314	12.706	31.821	63.657

图1：一个自由度下 Student*t*分布的临界值（来源：peo- ple.richland.edu）

译者。那么，这一评估结果的信度如何？

QS1 和 QS2 的样本均值 x 为 ($QS1 + QS2)/2 = 79.42$ ，这是基于两个测量值的样本。由两个数值组成的样本标准差 s 为$\sqrt{(QS1 - \text{x})^2 + (QS2 - \text{x})^2} = \sqrt{6.6049 \times 2}$，结果为 3.6345。

置信区间取决于所需的置信水平，而置信水平又取决于所翻译内容所属的主题领域。对于大多数字段，置信水平应至少为 80%。该置信水平对应的临界值 $t_{\alpha/2}$ 为 0.1（即 20% 除以 2，对应图形的一条尾部， $\alpha = 0.2$）；对于两个测量值（一个自由度，自由度 =1），临界值为 3.078，如图1和图2所示。

因此，这些测量值的误差范围为：

$$E = \frac{3.078 \times 3.6345}{\sqrt{2}} = 7.91 \qquad (12)$$

这意味着，这两个测量值的置信区间为 7. 91×2=15.8，这表明在 80% 的置信度下，真实质量分数位于以下区间内： [79.42 – 7.91,79.42 + 7.91], 或 [71.51,87.33]。

从这一结果可以看出，基于第二次测量，与单次测量相比，我们可以显著改善置信区间。我们获得的两种不同判断（观测值）将方差从单次判断的 25%（第 3 节）降至 9.96%（7.91/79.42），即在 80% 置信度设置下，区间宽度缩小超过两倍。

然而，与前一个示例一样，这个置信区间仍然相对较大。我们能否判断译者是否达到 89% 的通过/不通过阈值？答案是：由于样本均值低于 80%，且为 79.42，因此评估结果属于临界失败。

理想情况下，我们需要进行第三次测量，甚至获取更多观测值，以进一步缩小这个区间；但在生产环境中，受此类流程所需的成本和时间限制，额外的数据点很少会通过重复评估获得。

好消息是，出于翻译质量评估（TQE）目的，我们已经获得了可靠的信息：这是一次临界失败，因为样本均值低于阈值，因此一半以上的可能值低于通过阈值。对于像人类对质量的感知这样微妙、近乎无形的对象而言，这样的测量并不差。但只有基于对不同内容片段进行多次评估，才能获得性能历史记录并应用数据科学方法。

5 结论与未来工作

在评估翻译质量时，仅仅具备测量能力还不够；我们还需要了解测量的信度。自动评估质量时，即使重复进行多次，也会迅速得出相同的分数，这造成了一种精确度很高的错觉。遗憾的是，自动质量测量的结果不仅取决于语言对、机器学习系统、解码器和内容类型，也会因数据集而异，具体取决于数据的清理和格式化方式。鉴于这些因素，自动测量可以非常快速且“可靠”，但它可能（而且往往确实）无效，同时也缺乏一致性。目前，人类翻译质量评估（TQE）是获得有效的人类对质量的感知测量值的唯一途径，也被视为翻译质量评估的黄金标准。然而，即使评估过程完全正确，也应评估人工测量的评估者间信度（IRR）。即使评估者是经验丰富的语言学家，并接受过按适当体系进行评估的培训，

且客户要求也定义明确，评估者仍会因人类对质量的感知本质上取决于个人感知，而给出彼此接近但并不完全相同的评估结果。现实中的生产环境往往没有时间或预算，即使有第二位审校员，也无法验证翻译质量测量；而即便有第二位审校员，此类测量的评估者间信度较低，也会使确认首次测量结果变得困难。

本文首先研究了黄金标准人工质量测量的典型生产场景：翻译质量评估由一位经验丰富、受过培训的语言学家依据明确定义的客户规范进行，仅产生一个测量值，并据此对该测量值的信度作出结论。随后，我们以两位不同评审人员所做的两个测量值的 *Student's t* 分布分析为例，说明了研究结果。

根据第一项和第二项实验，我们可以得出结论：单次测量的信度很低，只能作为指示性参考。单次测量的置信区间（如第 3 节所示）宽达 25%，因此不能将一次评估作为流程决策的依据；还需要进行更多测量。例如，第二次测量可以缩小这一置信区间，使其缩小一半，达到约 10%（第 4 节）。

不过，我们可以说，置信区间的中点位于先前的平均结果与较低的近期测量值之间的中间位置。由此得出一条实用经验法则：**如果单次测量值偏离平均值，那么真实值很可能位于平均值与这次新测量值的中间**。因此，我们建议：不要对异常偏低的单次新评分反应过度，而应在先前的平均值和新评分之间取一个折中值，并将其视为最可能的结果。

第二次测量可能会显著缩小置信区间，但除仲裁期间外，很少进行第二次测量。因此，在翻译流程中，从对不同样本进行的其他评估中获取额外数据点，更为实际。

后续评估实际上被归入两类：（a）大多为通过，仅偶尔出现不通过；（b）其他所有情况（大多不通过，或有许多次不通过）。这种策略源于确保系统稳定地高于通过/不通过阈值的需求，从而保证结果质量。多次评估还能确认质量测量的有效性，并使人们能够应用正态分布统计学中广为人知的数学方法。

不过，值得注意的是，分析质量评估数据集需要采用适当的数据分析方法，例如：

- 剔除由无关因素导致的离群值。
- 对极小或极大的样本进行评估。
- 由于未正确应用指标而导致的错误评估，例如，重复错误计数不当。
- 由未经培训、带有主观倾向或心中设想了不同客户需求的评审人员所做的评估。

请记住，只有正确清理数据，数据科学才能获得良好的结果。使用未经清理的数据，可能会得出错误、有偏差、校准不当或不精确的结论和推断。

局限性

在这项工作中，我们讨论了如何在只有来自标注员（如翻译质量评估员）的一个或两个评估分数时，计算置信区间和评估可靠性。对于第一种情况，即只有一个新的观测分数时，我们假设已有预先估计／预期的分数可供使用，也就是用于 ARF 区间的分数。然而，在某些情况下未必如此；例如，对于缺乏先验知识的新设任务，获取该数值可能需要花费时间和金钱。在所考虑的第二种情况下，我们介绍了 Student's *t*-分布方法，并提供两个用于估计的人工判断分数。我们预计，这对观测数量较少的情况会有所帮助；不过，该方法确实需要借助引导公式和参数表进行一些数学计算，对于统计知识不多的译者或项目经理而言，可能无法立即方便地使用，而且无论如何都需要由受过教育的 AI 研究人员手动计算，

而且无论如何都需要由受过教育的 AI 研究人员手动计算。对于实际应用，可能需要进行初步设置，并为从业者提供额外的清晰简明指导。

伦理声明

本研究不存在伦理方面的顾虑，因为本文仅介绍用于计算人工评估置信度和信度的替代方法。

致谢

我们感谢 Whale 项目与我们分享其真实的翻译质量评估示例，以便我们应用所讨论的方法。我们感谢普渡大学统计学系的 Leonore Findsen 博士，以及里奇兰社区学院数学系的 James Jones 教授提供开放源码的讲义。我们感谢匿名评审人员对本研究提出的宝贵意见。本研究部分由 EP/V047949/1 号资助项目“将医院门诊信件整合到医疗数据空间”（资助方：英国研究与创新署／工程与自然科学研究委员会）资助。

参考文献

JH Abbott 和 JI Rosenblatt。1962。第一阶段仅有一次观测时的两阶段估计。统计数学年刊，14(1):229–235。

Alexandra Alekseeva、Serge Gladkoff、Irina Sorokina 和 Lifeng Han。2021 年。翻译质量评估（TQE）和译后编辑距离（PED）测量中置信区间的蒙特卡罗建模。载于 *Metrics 2021*：计量信息学与科学计量学研究研讨会（*SIG-MET*），*2021* 年 *10* 月 *23—24* 日。信息科学与技术协会。

Haifa Alrdahi、Uli Sattler 和 Goran Nenadic。2020 年。一种用于从寄生虫学文献中自动回答清单问题的文本挖掘模型。载于 *2020* 年国际计算与信息技术会议（*ICCIT-1441*），第 1—5 页。

Stefanie Beck、Bjarne Ruhnke、Malte Issleib、Anne Daubmann、Sigrid Harendza 和 Christian Z¨ollner。2016 年。专业人员、医学生和受训中小学生作为基础生命支持技能评估者时的评估者间信度分析。*BMC* 医学教育，16(1)：1—8。

Jyoti Belur、Lisa Tompson、Amy Thornton 和 Mi-randa Simon。2021。系统综述方法中的评分者间信度：探索差异与编码者决策。《社会学方法与研究》，50(2)：837–865。Manik Bhandari、Pranav Narayan Gour、Atabak Ash-faq、Pengfei Liu 和 Graham Neubig。2020。重新评估文本摘要中的评价。在*2020* 年自然语言处理实证方法会议论文集（*EMNLP*）中，第 9347–9359 页，线上版。计算语言学协会。R Clifford Blair 和 James J Higgins。1980。不同非正态分布下 Wilcoxon 秩和统计量与 Student t 统计量的检验效能比较。《教育统计学杂志》，5(4)：309–335。Alan B Cantor。1996 年。Cohen κ 的样本量计算。《心理学方法》，1(2)：150。

Parthena Charalampidou 和 Serge Gladkoff。2022。一个在 EMT课堂中应用新型机器翻译质量人工评估指标的案例。载于翻译与技术新趋势（*NeTTT*）会议，第 161–165 页。

Smiley W Cheng 和 James C Fu。1983。一种求得 t、 $\chi 2$和 f 分布临界值的算法。《统计与概率快报》，1(5):223–227。

Jacob Cohen。1960。名义量表的一致性系数。*Educational and Psychological Measurement*，20(1):37–46。

Rosario Delgado 和 Xavier-Andoni Tibau。2019。为什么应避免将 Cohen κ系数用作分类的性能测量指标。《公共科学图书馆·综合》，14(9):e0222916。

Markus Freitag、George Foster、David Grangier、Viresh Ratnakar、Qijun Tan 和 Wolfgang Macherey。2021。专家、错误与语境：一项关于机器翻译人工评估的大规模研究。*arXiv* 电子印本，第 arXiv:2104.14478 页。

George M Furnival、Timothy G Gregoire 和 Harry T Valentine。1989年。单次试验的置信区间和显著性检验。《统计学通讯：理论与方法》，18(10)：3749–3761。

Serge Gladkoff 和 Lifeng Han。2022。HOPE：一种以任务为导向、以人为中心的评估框架，通过专业译后编辑实现更有效的机器翻译评估。载于第十三届语言资源与评估会议论文集，第13–21页，马赛，法国。欧洲语言资源协会。

Serge Gladkoff、Irina Sorokina、Lifeng Han 和 Alexandra Alekseeva。2022。衡量翻译质量评估（TQE）中的不确定性。载于第十三届语言资源与评估会议论文集，第1454–1461页，法国马赛。欧洲语言资源协会。

YvetteGraham、TimothyBaldwin、Alistair Moffat 和 Justin Zobel。2017。机器翻译系统能否仅通过众包进行评估？自然语言工程，23(1):3–30。

Kevin A Hallgren。2012。计算观察数据的评审者间信度：概述与教程。心理学定量方法教程，8(1)：23。

Lifeng Han。2022a。机器翻译中的多词表达研究。博士论文，都柏林城市大学。

Lifeng Han。2022b。机器翻译评估概述。*arXiv* 预印本 *arXiv:2202.11027*。

Lifeng Han 和 Serge Gladkoff。2022。翻译评估方法的元评估：系统且最新的综述。载于 *LREC2022* 研讨会，马赛，法国。

Lifeng Han、Gareth Jones 和 Alan Smeaton。2020。AlphaMWE：构建带有多词表达标注的多语言平行语料库。载于多词表达与电子词典联合研讨会论文集，第44–57页，线上举行。计算语言学协会。

Lifeng Han、Alan Smeaton 和 Gareth Jones。2021。翻译质量评估：关于人工和自动方法的简要综述。载于第一届翻译建模研讨会：数字时代的翻译学论文集，第 15–33 页，在线版。计算语言学协会。Naomi D Harvey。2021。动物行为研究中评估者间信度、评估者内信度和重测信度简明指南。Michael Krauthammer 和 Goran Nenadic。2004。生物医学文献中的术语识别。《生物医学信息学杂志》，37(6)：512–526。Klaus Krippendorff。1987。关联、一致性与公平。《质量与数量》，21(2)：109–123。Klaus Krippendorff。2011。计算 Krippendorff α 信度。Seemab Latif、Mary McGee Wood 和 Goran Nenadic。2009。人工评估作文与 ROUGE 评估作文摘要之间的相关性。载于*2009* 年第八届自然语言处理国际研讨会，第 122–127 页。IEEE。Arle Lommel、Aljoscha Burchardt 和 Hans Uszkoreit。2014。多维质量指标（MQM）：用于声明和描述翻译质量指标的框架。*Tradum`atica: tecnologies delatraducci´o*，0：455–463。KHALED MILAD。2022。阿拉伯语与英语互译中的翻译记忆（TM）和机器翻译（MT）系统比较评估。载于翻译与技术新趋势（*NeTTT*）会议，第 142–151 页。

Nikola Milo˘sevi´c、Dimitar Marinov、Abdullah G¨ok 和 Goran Nenadi´c。2019。从网络爬取文本到项目描述：社会创新项目的自动摘要。载于自然语言在信息系统中的应用国际会议，第 157–169 页。施普林格。

Goran Nenadic、Sophia Ananiadou 和 John Mc- Naught。2004年。通过识别变体增强自动术语识别。载于*COLING 2004*：第*20*届计算语言学国际会议论文集，第604–610页。

CC Rodriguez。1996年。基于单次观测得出的置信区间。载于最大熵与贝叶斯方法：剑桥，英格兰，*1994*年，第十四届最大熵与贝叶斯方法国际研讨会论文集，第175–182页。施普林格。

Student。1908年。均值的可能误差。*Biometrika*，6(1):1–25。

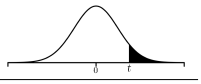
Harry T Valentine、George M Furnival 和 Timothy G Gregoire。1991年。森林研究中基于单次观测得出的置信区间。*Forest science*，37(1):370–373。

Cindy M Walker 和 Sakine G¨oc¸er S¸ahin。2020年。使用项目功能差异检验建构反应题的评估者间信度。*Educationaland Psychological Measurement*，80(4):808–820。Ka Wong、Praveen Paritosh 和 Lora Aroyo。2021年。跨重复实验信度——一种解读评估者间信度的实证方法。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第7053–7065页，线上。计算语言学协会。Yuping Wu、Lifeng Han、Valerio Antonini 和 Goran Nenadic。2022年。关于用于临床文本挖掘的跨领域预训练语言模型：在数据受限的微调中表现如何？载于*arXiv:2210.12770 [cs.CL]*。Hui Yang、Irena Spasic、John A Keane 和 Goran Nenadic。2009年。运用文本挖掘从临床出院摘要预测疾病状态。美国医学信息学会杂志，16(4):596–600。Xi Yang、Chengkun Wu、Goran Nenadic、Wei Wang 和 Kai Lu。2021年。从文献中挖掘卒中知识图谱。*BMC Bioinformatics*，22(10):1–19。

附录

普渡大学统计系的图2展示了 Student's*t* 分布的详细临界值。

Student's t分布的临界值。



	上尾概率：Pr($T > t$)									
df	0.2	0.1	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
1	1.376	3.078	6.314	7.916	10.579	12.706	15.895	31.821	63.657	636.619
2	1.061	1.886	2.920	3.320	3.896	4.303	4.849	6.965	9.925	31.599
3	0.978	1.638	2.353	2.605	2.951	3.182	3.482	4.541	5.841	12.924
4	0.941	1.533	2.132	2.333	2.601	2.776	2.999	3.747	4.604	8.610
5	0.920	1.476	2.015	2.191	2.422	2.571	2.757	3.365	4.032	6.869
6	0.906	1.440	1.943	2.104	2.313	2.447	2.612	3.143	3.707	5.959
7	0.896	1.415	1.895	2.046	2.241	2.365	2.517	2.998	3.499	5.408
8	0.889	1.397	1.860	2.004	2.189	2.306	2.449	2.896	3.355	5.041
9	0.883	1.383	1.833	1.973	2.150	2.262	2.398	2.821	3.250	4.781
10	0.879	1.372	1.812	1.948	2.120	2.228	2.359	2.764	3.169	4.587
11	0.876	1.363	1.796	1.928	2.096	2.201	2.328	2.718	3.106	4.437
12	0.873	1.356	1.782	1.912	2.076	2.179	2.303	2.681	3.055	4.318
13	0.870	1.350	1.771	1.899	2.060	2.160	2.282	2.650	3.012	4.221
14	0.868	1.345	1.761	1.887	2.046	2.145	2.264	2.624	2.977	4.140
15	0.866	1.341	1.753	1.878	2.034	2.131	2.249	2.602	2.947	4.073
16	0.865	1.337	1.746	1.869	2.024	2.120	2.235	2.583	2.921	4.015
17	0.863	1.333	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.965
18	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922
19	0.861	1.328	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.883
20	0.860	1.325	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.850
21	0.859	1.323	1.721	1.840	1.988	2.080	2.189	2.518	2.831	3.819
22	0.858	1.321	1.717	1.835	1.983	2.074	2.183	2.508	2.819	3.792
23	0.858	1.319	1.714	1.832	1.978	2.069	2.177	2.500	2.807	3.768
24	0.857	1.318	1.711	1.828	1.974	2.064	2.172	2.492	2.797	3.745
25	0.856	1.316	1.708	1.825	1.970	2.060	2.167	2.485	2.787	3.725
26	0.856	1.315	1.706	1.822	1.967	2.056	2.162	2.479	2.779	3.707
27	0.855	1.314	1.703	1.819	1.963	2.052	2.158	2.473	2.771	3.690
28	0.855	1.313	1.701	1.817	1.960	2.048	2.154	2.467	2.763	3.674
29	0.854	1.311	1.699	1.814	1.957	2.045	2.150	2.462	2.756	3.659
30	0.854	1.310	1.697	1.812	1.955	2.042	2.147	2.457	2.750	3.646
31	0.853	1.309	1.696	1.810	1.952	2.040	2.144	2.453	2.744	3.633
32	0.853	1.309	1.694	1.808	1.950	2.037	2.141	2.449	2.738	3.622
33	0.853	1.308	1.692	1.806	1.948	2.035	2.138	2.445	2.733	3.611
34	0.852	1.307	1.691	1.805	1.946	2.032	2.136	2.441	2.728	3.601
35	0.852	1.306	1.690	1.803	1.944	2.030	2.133	2.438	2.724	3.591
36	0.852	1.306	1.688	1.802	1.942	2.028	2.131	2.434	2.719	3.582
37	0.851	1.305	1.687	1.800	1.940	2.026	2.129	2.431	2.715	3.574
38	0.851	1.304	1.686	1.799	1.939	2.024	2.127	2.429	2.712	3.566
39	0.851	1.304	1.685	1.798	1.937	2.023	2.125	2.426	2.708	3.558
40	0.851	1.303	1.684	1.796	1.936	2.021	2.123	2.423	2.704	3.551
41	0.850	1.303	1.683	1.795	1.934	2.020	2.121	2.421	2.701	3.544
42	0.850	1.302	1.682	1.794	1.933	2.018	2.120	2.418	2.698	3.538
43	0.850	1.302	1.681	1.793	1.932	2.017	2.118	2.416	2.695	3.532
44	0.850	1.301	1.680	1.792	1.931	2.015	2.116	2.414	2.692	3.526
45	0.850	1.301	1.679	1.791	1.929	2.014	2.115	2.412	2.690	3.520
46	0.850	1.300	1.679	1.790	1.928	2.013	2.114	2.410	2.687	3.515
47	0.849	1.300	1.678	1.789	1.927	2.012	2.112	2.408	2.685	3.510
48	0.849	1.299	1.677	1.789	1.926	2.011	2.111	2.407	2.682	3.505
49	0.849	1.299	1.677	1.788	1.925	2.010	2.110	2.405	2.680	3.500
50	0.849	1.299	1.676	1.787	1.924	2.009	2.109	2.403	2.678	3.496
60	0.848	1.296	1.671	1.781	1.917	2.000	2.099	2.390	2.660	3.460
70	0.847	1.294	1.667	1.776	1.912	1.994	2.093	2.381	2.648	3.435
80	0.846	1.292	1.664	1.773	1.908	1.990	2.088	2.374	2.639	3.416
90	0.846	1.291	1.662	1.771	1.905	1.987	2.084	2.368	2.632	3.402
100	0.845	1.290	1.660	1.769	1.902	1.984	2.081	2.364	2.626	3.390
120	0.845	1.289	1.658	1.766	1.899	1.980	2.076	2.358	2.617	3.373
140	0.844	1.288	1.656	1.763	1.896	1.977	2.073	2.353	2.611	3.361
180	0.844	1.286	1.653	1.761	1.893	1.973	2.069	2.347	2.603	3.345
200	0.843	1.286	1.653	1.760	1.892	1.972	2.067	2.345	2.601	3.340
500	0.842	1.283	1.648	1.754	1.885	1.965	2.059	2.334	2.586	3.310
1000	0.842	1.282	1.646	1.752	1.883	1.962	2.056	2.330	2.581	3.300
∞	0.842	1.282	1.645	1.751	1.881	1.960	2.054	2.326	2.576	3.291
	60%	80%	90%	92%	94%	95%	96%	98%	99%	99.9%
	置信水平									

注： $t(\infty)_{\alpha/2} = Z_{\alpha/2}$ 是我们的记法。

图2：Student$t$分布的临界值列表（来源：stat.purdue.edu）