ICS 03.080.99
CCS A 20

团 体 标 准

T/CAQ XXXXX—2023

机器翻译服务质量评价规范—中英双向

Specification for Quality Evaluation of Machine Translation Service —English-Chinese

（征求意见稿）

2023 - XX - XX 发布 2023 - XX - XX 实施

中 国 质 量 协 会 发 布

目　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利，本文件的发布机构不承担识别这些专利的责任。

本文件由华为技术有限公司和中国质量协会联合提出。

本文件由中国质量协会归口。

本文件起草单位：华为技术有限公司、中国质量协会数字化分会、沈阳雅译网络技术有限公司、腾讯技术有限公司、南京大学、小米科技责任有限公司、阿里巴巴（中国）网络技术有限公司、北京字节跳动科技有限公司、中质协质量保证中心、。

本文件主要起草人：XXX,XXXXX……

引　言

随着深度学习和计算机算力的发展，机器翻译已经成为人工智能最成功、最广泛的应用之一，机器翻译已应用于多行业、多领域、多场景。

机器翻译依靠如数据稀疏建模、篇章模型、预训练模型等实现语言建模能力和知识推理能力，具备服务质量一致性和可储存性的特征。同时，机器翻译的使用者对其的关注涉及忠实度、流利度、兼容性、对预期用途或规定条件的适用性等特性，因此机器翻译服务提供给顾客的是一种兼具实物载体、算法（能力）和服务的商品（产品）。

以统一的规范，对机器翻译服务的质量进行科学、有效、高效地评价，证实产品与顾客要求的符合性，对于更好地应用机器翻译服务尤为重要。

本文件基于国内知名机器翻译提供商的内部质量要求和业内通用测试要求，统一和细化了评价方法、评价过程、评价能力要求和评价的具体指标和判断规则，同时明确了机器翻译有关术语，填补国家标准和行业标准的空白。

翻译语向众多，不同语系之间存在较大差异，本文件聚焦使用最为广泛的中英双向机器翻译的质量评价，为其他语向机器翻译的质量评价提供有益借鉴。

机器翻译服务质量评价规范—中英双向

1 范围

本文件规定了中英双向机器翻译服务质量评价的方式、方法、评价标准和评价能力要求。

本文件的应用场景包括但不限于机器翻译服务上线、国际和国内机器翻译比赛、机器翻译服务竞品对比、机器翻译模型上线验证、机器翻译模型训练验证等。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件。不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 19363.1—2022 翻译服务　第 1 部分：服务要求

GB/T 19682—2005　翻译服务译文质量要求

GB/T 19000　质量管理体系 基础和术语（idt ISO 9000：2000）

3 术语和定义

下列术语和定义适用于本文件。

3.1

机器翻译服务 machine translation service

用计算机程序将文本或语言从一种自然语言转换成另一种自然语言的服务。

3.2

机器翻译质量 machine translation quality

机器翻译译文在传达原文语义，表达流畅等方面，帮助服务对象进行信息阅读、理解和交流的程度。

3.3

质量评价 quality evaluation

采用特定评价方法和指标衡量机器翻译质量的过程。

3.4

人工质量评价 human evaluation

评价人员根据预先设定的标准，对机器翻译译文的忠实度、流利度等方面进行评分或提供反馈意见。

机器翻译的人工评价方式包含但不限于以下几种：人工打分评估、错误分类评估等。

3.5

自动质量评价 automatic evaluation

有参考译文的情况下，采用评估指标对机器翻译译文进行评价。常用的评估指标包括但不限于：BLEU、COMET。

3.6

测试集 test set

用于评估机器翻译模型质量的一组数据样本的集合，包含原文（必选）和对应的人工译文（可选）。

3.7

参考译文 reference translation

测试集的人工译文部分，该部分可作为参考答案用于机器翻译质量评估。

3.8

错误分类 error typology

人工评价过程中，评价人员标注机器翻译错误时使用的分类信息。

3.9

忠实度 fidelity

用来描述译文与原文之间的忠实程度，即机器翻译译文与原文的语义之间的一致性。

3.10

流利度 fluency

用来描述译文表达的流畅程度，流利度衡量了机器翻译译文在语言上的可读性和可理解性。

4 机器翻译服务质量评价方式

4.1 自动评价

本文件采用如下自动评价指标衡量机器翻译质量：

——SacreBLEU

——COMET20

工具下载地址参照附录 A。

4.2 人工评价

4.2.1 直接评分法

本文件采用直接评分法评价机器翻译质量，评分要求及方法如下：

——提供原文、参考集（可选）和待评价机器翻译的译文；

——基于公平性考虑，采用盲测，即译文来源等信息不对评价人员公开；

——打分方法为评价人员判断待测机器翻译的译文还原了原文信息的程度，并据此给出相应的分数，得分区间为[0,100]，扣分最小颗粒度为 1 分。

得分区间和对应的评价标准参照附录 B。

4.2.2 错误标注

可视实际需求确定是否进行错误标注。本文件将机器翻译错误分为忠实度错误和流利度错误 2 个大类和 10 个子类。具体错误分类参照附录 C。

5 机器翻译服务质量评价过程

5.1 确定评价范围和方法

基于业务需求确定测试集和评价方法（自动评价、人工评价的直接评分法）。

5.2 构建测试集

5.2.1 构建自动评价测试集

自动评价测试集为中英双语的文本，需满足如下要求：

——规模：1000-5000 句（默认 2000 句）；
——句长：5-200 字/词之间，且 5-100 字/词的句子占比不少于 90%；
——语义：双语互译性高；句子完整且语义相对独立；句式、内容丰富；
——地道性：原文为本地母语；参考译文翻译正确，符合目标语言的表达习惯；
——领域：所有句子来源于待测领域。对于重点领域，确保其所有子领域覆盖全面且均衡。

5.2.2 构建人工评价测试集

人工评价测试集为中英双语文本，规模为 200-500 句（默认 500 句），其他要求同自动评价测试集。

5.3 确定评价人员

5.3.1 确定自动评价人员

采用自动评价时，评价人员需具备评价工具的操作执行能力。

5.3.2 确定人工评价人员

采用人工评价时，至少需要两名评价人员；如需错误标注，至少需其中一名人员完成标注。评价人员需具备如下能力及资质要求：

a） 语言及翻译能力

评价人员需要具备熟练的中文、英文语言能力，有被认可的外语水平证书或与之相当的证书，有两年以上的机器翻译评测经验或机器翻译译后编辑经验，能够准确识别机器翻译文本中的语法、语义、文化差异、一致性等问题，能够运用符合中英文化特征的行为规范、价值体系以及区域特性等相关信息的能力。

b） 领域专业知识

评价人员需要具备高效拓展专业知识的能力，能够准确理解该领域术语和语言风格。

c） 技术能力

评价人员能够利用技术资源，包括使用评价工具和信息技术系统来支撑评价过程。

5.4 执行评价

5.4.1 执行自动评价

自动评价人员执行自动评价工具输出机器翻译质量得分。相关工具下载路径，参见附录 A。

5.4.2 执行人工评价

a） 完成评价前准备。

评价前应做好以下工作：
——了解评价范围和方法；
——熟悉所涉领域的知识；
——查阅单词和专业术语；
——如有需要，和需求方确认专业和术语上的问题。

b） 进行评价并输出评价报告。

评价人员仔细阅读并比对原文和机器翻译文本，按照统一的标准对每句机器翻译译文给出一个得分。每句待测译文的最终得分取所有评价人员评分的平均值，模型的最终得分为所有待测译文最终得分的平均值。

可结合实际需求，确定是否进行错误标注。若由一人进行错误标注，直接汇总每一类错误数；如有

多人参与错误标注，先汇总每人标注的每一类错误数，再取每一类错误数的平均值。

此外，评价人员可提供总体反馈和改进建议，并记录评价过程中发现的典型案例供后续分析使用。

6 机器翻译服务质量评价能力要求

机器翻译服务质量评价提供方需具备如下能力：

——测试集：能够基于需求构建测试集，包括但不限于数据提取、清洗、筛选等；

——评价标准：具有明确、可按需定制的评价标准，确保能够有效、可靠地评价机器翻译质量；

——专业的评价人员：评价人员拥有从业经验，能够基于评价标准提供客观一致的评价结果；

——工具或平台：具有工具或平台，能够有效支撑数据管理和翻译质量评价服务流程。

7 机器翻译服务准入基线及测试集

7.1 机器翻译服务准入基线

7.1.1 自动评价方式下的准入基线

a) 通用领域准入基线

通用领域下，本标准定义如下机器翻译服务准入基线：

语向	测试集	SacreBLEU	COMET20
中到英	WMT22	25.00	32.00
	CCMT21	33.00	36.00
英到中	WMT22	48.00	59.00
	CCMT21	48.00	52.00

注 1： 本标准采用 WMT22、CCMT21 数据作为通用开发集，定义以上机器翻译服务准入基线。

注 2： 机器翻译服务提供方可以从 WMT、CCMT 官方网站，获取公开数据，进行机器翻译服务质量自测、质量提升等相关活动。

b) 垂直领域准入基线

因领域不固定，无法设置统一基线，可与行业开源的机器翻译系统进行横向对比，了解质量优劣，确定是否满足准入基线。

7.1.2 人工评价方式下的准入基线

人工评价采用百分制，直接评分大于等于 80 分，作为机器翻译服务准入基线，不区分领域。

7.2 测试集

机器翻译服务认证活动中使用的测试集为非公开数据，符合 5.2 有关要求，由第三方公正性平台提供或确认。

附 录 A
（规范性）
自动评价工具下载地址

SacreBLEU：https://github.com/mjpost/sacrebleu
COMET：https://unbabel.github.io/COMET/html/faqs.html#which-comet-model-should-i-use

附 录 B
（规范性）
直接评分法得分区间

表 B.1 列出了采用直接评分法的得分区间和对应的评价标准。

表 B.1 得分区间和对应的评价标准

得分区间	标准
0-20	译文语义不明或完全错误，只有小部分字、短语正确且可读性极差，难以理解。
21-40	译文与原文极少部分语义相同但关键信息缺失或错误且可读性较差，大量不地道、不流利表达和语法错误。
41-60	译文能体现部分关键语义，但大量非关键语义错误且流利度、地道性欠佳。
61-80	译文基本能传达原文关键语义，但存在部分非关键信息错误，同时存在语法错误和非地道性表达。
81-100	译文可呈现原文语义，只存在少量非关键信息错误且表达较地道流畅。

附 录 C
（规范性）
机器翻译错误分类

表 C.1 规定了机器翻译错误类别。

表 C.1 机器翻译错误分类

错误大类	错误子类
忠实度	术语/命名实体错误
	错译
	漏译（原文中内容在译文中未体现）
	过译
	未译（原文内容直接搬到译文中，未翻译）
流利度	语域/风格错误
	拼写错误
	标点错误
	语法错误
	晦涩拗口

参 考 文 献

[1] Freitag M, Foster G, Grangier D, et al. Experts, errors, and context: A large-scale study of human evaluation for machine translation[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 1460-1474.

[2] Kocmi T, Bawden R, Bojar O, et al. Findings of the 2022 conference on machine translation (WMT22)[C]//Proceedings of the Seventh Conference on Machine Translation (WMT). 2022: 1-45.

[3] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.

[4] Rei R, Stewart C, Farinha A C, et al. COMET: A neural framework for MT evaluation[J]. arXiv preprint arXiv:2009.09025, 2020.