ICS 35.020
L 70

团　　体　　标　　准

T/GDCSA XXX-XXXX

信息技术 通话智能翻译产品质量评测方法

Information technology—Quality evaluation method for call intelligent translation products

（征求意见稿）

2024-XX-XX 发布　　2024-XX-XX 实施

广东省网络空间安全协会　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。

本文件由中移互联网有限公司提出。

本文件由广东省网络空间安全协会归口管理。

本文件起草单位：。

本文件主要起草人：。

信息技术 通话智能翻译产品质量评测方法

1 范围

本文件规定了通话过程中可实时机器翻译,并以字幕形式呈现结果的通话智能翻译产品的指标要求与评测方法，包括评测环境、流程、指标项与评测方法。

本文件适用于通话智能翻译类产品的质量评估。

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 21023-2007 中文语音识别系统通用技术规范

GB/T 25000.51-2016 系统与软件工程 系统与软件质量要求和评价（SQuaRE）第51部分：就绪可用软件产品（RUSP）的质量要求和测试细则

GB/T 41867-2022 信息技术 人工智能 术语

YD/T 4394.3-2023 自然语言处理技术及产品评估方法 第 3 部分：智能翻译机翻译服务译文质量要求

ISO 17100-2015 翻译服务 翻译服务要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能 AI artificial intelligence

人工智能系统相关机制和应用的研究和开发。

[来源：GB/T 41867-2022]

3.2

语音识别 speech recognition

将人类的声音信号转化为文字或者指令的技术。

[来源：GB/T 21023-2007]

3.3

机器翻译 MT machine translation

使用计算机系统将文本或语音从一种自然语言自动翻译为另一种自然语言。

[来源：ISO 17100-2015]

3.4

通话智能翻译 call intelligent translation

使用机器翻译技术，将通话对方的声音从一种自然语言自动翻译为目标自然语言，并以字幕形式呈现翻译结果。

3.5

语料 corpus

作为通话智能翻译业务的语音物料，具体为一定时长的语音文件。

3.6

关键字词 key word

对译文理解和使用结果的正确与否产生直接作用的字词。

[来源：GB/T 19682-2005]

3.7

语义差错 semantic error

由于对原文理解的错误，导致译文表述的错误。

[来源：GB/T 19682-2005]

3.8

有效性 effectiveness

用户实现指定目标的准确性和完备性。

[来源：GB/T 25000.10-2016]

3.9

易学性 learnability

在指定的使用周境中，产品或系统在有效性、效率、抗风险和满意度特性方面为了学习使用该产品或系统这一指定的目标可为指定用户使用的程度。

[来源：GB/T 25000.10-2016]

3.10

测试终端 test terminal

为测试通话智能翻译业务所需的手机或电脑载体。

3.11

主叫 caller

使用通话智能翻译业务时，发起通话一方的测试终端。

3.12

被叫 callee

使用通话智能翻译业务时，接收通话一方的测试终端。

4 评测环境要求

4.1 评测要素

4.1.1 所需硬件

所需硬件应包含：

a） 音频播放设备。

b） 支持视频通话功能的两台智能手机。

c） 已开通翻译业务的测试号卡。

d） 一台电脑设备。

4.1.2 所需软件

所需软件应包含：

a） 具备图像帧数截取功能的视频播放软件。

b） 手机端录屏软件，可以同时录制视频和音频。

4.1.3 所需素材

所需素材应包含：

a） 准备产品支持语种的测试语料音频文件，要求官方标准发音，语料播音清晰，无明显杂音，语料长度为5分钟，语速均衡，语料内容应无自创词，语义应可被大众理解。

b） 语料内容包括外贸、跨境电商、国际会议安排、留学、旅游、日常沟通等场景，覆盖尽可能多的语法表达。

c） 上述测试语料的文字原始文本，要求文本字句与语料一一对应，且无语法错误，字句词语没有歧义。

4.2 评测前物理环境准备

评测前物理环境准备工作应包含：

a） 声音环境：确保所处环境相对安静，需要在没有明显杂音和人声干扰的安静环境下，噪音音量不大于40分贝。

b） 网络信号：所处环境网络信号强度稳定在-80dBm以上。

c） 设备状态：确保测试号卡已开通翻译业务且测试终端能够正常收音。音频播放设备可正常放音，音量不低于60分贝。

4.3 评测前业务操作状态确认

评测前业务操作状态确认应包含：

a） 空间环境不干扰：当前处于安静的环境中，区分测试终端A、B分别为主叫和被叫，且A所在空间声音无法传至B处。

b） 终端可正常发起呼叫：测试终端A作为主叫，对测试终端B发起智能翻译通话呼叫，测试终端B应答呼叫。

c） 智能翻译功能已开启：音频播放设备对测试终端A播放测试语料，查看测试终端B确认智能翻译功能开启。

d） 录屏软件可正常工作：连接通话后，测试终端B打开屏幕录制工具，进行录制，结束通话与视频录制，确认视频画面与声音可正常播放。

5 基本测试流程

基本测试流程规定了测试人员的一般测试步骤、测试记录的操作规范，各项指标的评测方法应在基本测试流程要求的基础上开展，包括以下要求：

a） 准备语料播放设备：音频播放设备准备好待播放的语料，将音频播放设备放置到测试终端 A 的语音输入端口。

b） 开始录制视频：测试终端 B 打开录屏软件，开启录屏功能。

c） 拨通视频通话：测试终端 A 作为主叫，对测试终端 B 发起视频呼叫，测试终端 B 接听。

d） 播放测试语料：打开音频播放设备，开始播放测试语料。

e） 挂断电话并停止录屏：待语料播放完毕后，挂断电话并停止录屏。

f） 导出翻译文件：将录屏文件导入视频播放软件，查看机器翻译的文本，并通过人工将写文本转录至 word 文档，形成翻译文本。

6 指标及评测方法

6.1 评测指标体系

通话智能翻译产品质量评测指标体系见表1。

表 1 通话智能翻译产品质量评测指标体系

维度	指标	评测方法
翻译能力范围	支持语种数	语种数量
翻译功能启用	功能开关有效性	成功率
	功能开关易学性	主观评估
	自动识别语种	支持与否
翻译效果感知	语音识别准确率	字准确率
	译文可理解度	主观评分
	翻译即时性	时延
	字幕可读性	主观评估

6.2 翻译能力范围

6.2.1 支持语种数

评估产品支持的汉语、英语、日语、韩语、俄语、德语、西班牙语、意大利语等常见语种机器翻译的数量，方法如下：

a） 根据产品说明支持的语种，准备各语种双向的测试语料。

b） 基于技术验证的方式，依次输入测试语料，通过通话转译的文本及对应机器翻译结果，判定其是否支持被测语种。

6.3 翻译功能启用

6.3.1 功能开关有效性

评价通话过程中，用户是否可正常开启或关闭机器翻译功能，方法如下：

a） 通话过程中，测试终端 B 根据操作指引开启机器翻译功能。

b） 测试终端 A 持续输入语料。

c） 等待测试终端 B 通话界面出现机器翻译结果。

d） 测试终端 B 根据操作指引关闭机器翻译功能。

e） 等待测试终端 B 通话界面翻译结果消失。

f） 计算成功率见式（1）：

$$P=\text{成功开关总数}/\text{尝试开关总数}\times 100\% \quad (1)$$

式中：

P——成功率。

6.3.2 功能开关易学性

评价通话过程中，功能开启或关闭的操作方式、操作引导是否简单易懂，方法如下：

a） 通话过程中，测试终端 B 根据操作指引开启、关闭机器翻译功能。

b） 按照操作方式便捷好记、操作引导清晰的原则评价上述操作过程。

c） 记录测试过程中，违背上述指标原则的问题。

6.3.3 自动识别语种

基于产品支持的语种，评估进入通话时，是否支持自动识别需要翻译的语种，方法如下：

a） 根据产品支持的语种，准备各语种双向的测试语料。

b） 基于技术验证的方式，依次输入测试语料，通过通话转译的文本及对应机器翻译结果，判定其是否支持自动识别语种。

6.4 翻译效果感知

6.4.1 语音识别准确性

评价产品是否能够准确识别输入的语音，并将其转写为对应的文字，方法如下：

a） 测试终端 A 输入语料，播放时长 5 分钟，测试终端 B 显示转译字幕，进行屏幕录制。

b） 通过视频软件打开录制的视频，提取语音转写文本，人工将转写文本与语料原文本进行对比评测，核对多转、漏转、错转情况。

c） 计算语音转写文本的字准确率，计算方法见式（2）：

$$A=（1-（W_1+W_2+W_3）/W）\times 100\% \quad (2)$$

式中：

A——字准确率；

W_1——多转字数；

W_2——漏转字数；

W_3——错转字数；

W——产品语音转写文本的总字数。

6.4.2 译文可理解度

评价译文是否能够流畅、完整地反映原文语义，方法如下：

a） 测试终端 A 输入语料，播放时长 5 分钟，测试终端 B 显示转译字幕，进行屏幕录制。

b） 通过视频软件打开录制的视频，提取机器翻译文本，人工将机器翻译的译文与标准译文进行对比评测，按照 0-5 分打分，评分规则见表 2。

c） 以译文可理解度作为机器翻译效果的度量，计算方法见式（3）：

$$C=T_1/T\times 100\% \quad (3)$$

式中：

C——译文可理解度；

T_1——译文句子评分之和；

T——译文句子总数×5 分。

表 2 译文可理解度评分规则

译文评分	评分标准
1 分	译文不知所云、晦涩难懂，语义差错程度大，与原文表达的内容完全不相符。
2 分	译文有小部分关键字词和原文相符，但是存在严重漏译、错译、多译，或者逻辑顺序错误、严重语法错误、语义差错等问题。
3 分	译文大致表达了原文的意思，与原文有局部出入，可推测出基本语义，存在表达语义的关键字词翻译不当、非关键字词漏译或错译等。
4 分	译文基本传达了原文的意思，语序相对流畅，但是部分关键字词存在用词不当、搭配不得体问题。

5 分	译文准确表达原文语义，语法结构正确，语句流畅，可存在个别错译，但不影响整体的语义表达。

6.4.3 翻译即时性

评价接收到通话对方的语音后，显示为转译字幕所需时间，方法如下：

a） 测试终端 A 输入一段语料，测试终端 B 听到这段语料的末尾字时记录时间 T_1，可通过视频软件打开录制的视频，以音频波形无波动来判断末尾字播放结束。

b） 查看测试终端 B 屏幕末尾字转译结束时，记录时间 T_2。

c） 翻译时延 $T=T_2-T_1$。

d） 重复步骤 a-c，操作 30 次。

e） 计算平均翻译时延。

6.4.4 字幕可读性

评价通话过程中译文字幕是否容易阅读，方法如下：

a） 通话过程中开启机器翻译功能，测试终端 A 输入语料，通过测试终端 B 查看转译字幕效果。

b） 评价字体大小、颜色、标点符号、换行等呈现方式是否符合用户阅读习惯与感受。

c） 记录测试过程中，违背上述指标原则的问题。

参 考 文 献

[1] GB/T 19682-2005 翻译服务译文质量要求

[2] GB/T 21023-2007 中文语音识别系统通用技术规范

[3] GB/T 25000.51-2016 系统与软件工程 系统与软件质量要求和评价（SQuaRE）第51部分：就绪可用软件产品（RUSP）的质量要求和测试细则

[4] GB/T 41867-2022 信息技术 人工智能 术语

[5] YD/T 4394.3-2023 自然语言处理技术及产品评估方法 第3部分：智能翻译机

[6] ISO 17100-2015 翻译服务 翻译服务要求