ICS 35.240.01
L 77

中华人民共和国国家标准

GB/T XXXXX.2—XXXX

信息技术　智能语音交互测试方法 第2部分：语义理解

Information technology—Intelligent speech interaction testing method—Part 2: Semantic understanding

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家市场监督管理总局
中国国家标准化管理委员会　发布

目　　次

前言 III
引言 IV
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　概述 1
5　测试准备和执行 1
5.1　测试数据集 1
5.2　测试工具 2
5.3　测试环境 3
5.4　测试执行 3
5.5　结果判定 3
6　功能测试方法 3
6.1　意图理解 3
6.2　命名实体识别 3
6.3　敏感信息辨别 3
6.4　语义拒识 3
6.5　信息检索 4
6.6　文本相似度 4
6.7　文本修改 4
6.8　语义修正 4
6.9　自然语言生成 4
6.10　逻辑推理 5
6.11　对话引导和推荐 5
6.12　上下文相关的多轮会话 5
7　性能测试方法 5
7.1　理解效果 5
7.2　理解效率 8
7.3　系统稳定性 9
附　录　A　（资料性）　主观体验测试 10
A.1 概述 10
A.2 测试项 10
A.3 测试方法 10
参考文献 12

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。

GB/T XXXXX《信息技术　智能语音交互测试方法》为GB/T 36464《信息技术　智能语音交互系统》提供基础性测试方法。

本文件是GB/T XXXXX《信息技术　智能语音交互测试方法》的第2部分。GB/T XXXXX已经发布了以下部分：

——第1部分：语音识别；

——第2部分：语义理解。

本文件由全国信息技术标准化技术委员会（SAC/TC 28）提出并归口。

本文件起草单位：XXX。

本文件主要起草人：XXX。

引　言

智能语音交互在智能家居、智能客服、移动终端、车载终端等诸多领域应用广泛，已成为当前人机交互的重要方式之一。随着智能语音交互越来越深入到生产生活的方方面面，需要对智能语音交互的系统参考框架、基础技术要求、互联网接口要求等进行统一规范，在这方面，我国已制定了支撑智能语音交互系统的基础性国家标准。在此基础上，也需要用统一的测试方法和评价标准来对智能语音交互系统的能力进行评测，为智能语音交互相关的产品和服务提供评测的基础方法和依据。

智能语音交互包括语音识别、语义理解和语音合成三个基本环节，各环节所涉及的测试对象、测试项目、测试环境和测试方法均有所不同。GB/T XXXX《信息技术　智能语音交互测试方法》旨在确立和描述适用于智能语音交互各环节的测试项和测试方法，拟由三个部分构成。

——第 1 部分：语音识别。目的在于为智能语音交互中的语音识别环节提供测试项和测试方法。

——第 2 部分：语义理解。目的在于为智能语音交互中的语义理解环节提供测试项和测试方法。

——第 3 部分：语音合成。目的在于为智能语音交互中的语音合成环节提供测试项和测试方法。

信息技术　智能语音交互测试方法　第2部分：语义理解

1　范围

本文件描述了智能语音交互测试中语义理解（子）系统的测试项和测试方法。

本文件适用于智能语音交互测试中语义理解（子）系统测试的设计和实施。

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 36464.1　信息技术　智能语音交互系统　第1部分：通用规范

3　术语和定义

GB/T 36464.1界定的以及下列术语和定义适用于本文件。

3.1

命名实体　named entity

具有特指或唯一含义的指称名字的实体。

4　概述

智能语音交互测试中的语义理解测试主要包括功能测试与性能测试，具体如下：

——　功能测试用于检查被测系统是否提供了语义理解相关的各项功能，具体包括；

——　性能测试用于检查被测系统中语义理解相关的各项性能，具体包括理解效果，理解效率，系统稳定性。

测试过程中，根据被测系统技术规范进行在线/离线的功能和性能测试。

注1：本文件对所列出的功能测试项选择不做要求，实际测试时根据被测系统的功能要求和测试需求进行选择。

注2：语义理解测试包括客观测试和主观体验测试，主观体验测试项和测试方法见附录A。

5　测试准备和执行

5.1　测试数据集

在测试开始前，应通过人工编写或采集的方式制作测试数据集。可根据不同测试项划分出多个测试数据集，在实际测试时可根据需要选择测试数据集。测试数据集类型和要求符合表1和表2的要求。

表 1　测试文本类型和要求

序号	文本分类		数量
1	常用文本	单字、词语文本	不少于 5 条
2		短语文本	
3		单句文本	
4		对话文本	
5		段落文本	
6		文章文本	
7	特殊文本	敏感信息文本	不少于 1000 条
8		命名实体文本，如：人名、地名等，覆盖已定义业务相关命名实体	
9		特殊格式文本，如：数字、日期时间、英文大小写等	不少于 5 条
10		特定语种文本，如：中文、英文、韩文等	
11		特殊编码文本，如：utf-8、gbk 等	
12		特殊符号文本，如：包含逗号、句号、问号等符号文本	
13	异常文本	乱码文本：如 utf8 和 gbk 混合编码等	不少于 5 条
14		不支持语种文本	

表 2　测试数据集类型和要求

序号	测试数据分类		文本要求	数量
1	已定义场景/业务文本数据	已定义场景/业务一般文本数据	文本长度：数据较多情况下，统计文本长度分布，根据此分布来控制文本长度数量分布；否则根据常用文本长度平均值的正态分布，控制不同文本长度数量分布。 文本类型：符合表 1 的要求。	每个业务不少于 200 条
2		已定义场景/业务常用文本数据		每个业务至少覆盖 3 次，可持续收集
3	未定义场景/业务文本数据	同领域，未定义场景/业务一般文本数据		每个业务至少覆盖 3 次
4		同领域，未定义场景/业务常用文本数据		每个业务至少覆盖 3 次，可持续收集
5		闲聊	均不少于 5 个字符	不少于 1000 条
6		无意义、无逻辑文本数据	均不少于 5 个字符	不少于 100 条

5.2　测试工具

a)　可编程测试工具要求如下：

- 应能调用被测系统开放接口；
- 应能对工具配置文件进行定制；
- 应能接收文本数据并将其输入至被测系统；
- 应能进行功能测试及其相应的性能测试；
- 应能以文本形式获取被测系统运行结果。

b)　测试统计工具要求如下：

- 应能自动不同测试指标的系统运行结果进行统计和分析；

- 应能自动对系统运行结果和标准结果对比文件进行比对。

c) 资源监测工具应能监测内存、CPU、GPU、句柄数等系统资源指标。

5.3 测试环境

应根据被测系统的功能和性能要求，按照被测系统的使用场景进行软硬件环境配置。

5.4 测试执行

应使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果。

5.5 结果判定

应根据被测系统技术要求对系统在各测试项上的运行结果进行记录、分析和判定，形成测试结果。

6 功能测试方法

6.1 意图理解

测试内容：检查被测系统是否提供理解说话人的意图的功能，包括但不限于以下具体功能。

a) 模糊识别：能正确处理错别字、同义词、多字少字问题。

b) 语义抽取：能抽取语义要素和说话人关键意图，包括：

- 命名实体抽取，被测系统能自动对文本中表达关键意图的命名实体进行抽取；
- 关键词抽取，被测系统能自动对文本中表达意图的关键词进行抽取；
- 语义关系抽取，被测系统能自动对文本中语义关系三元组进行抽取。

c) 语义排序：能在语义理解结果中给出多个排序后的理解结果，供说话人进行选择或二次确认。

d) 意图分类：检查被测系统是否提供对说话人的关键意图进行预测，将输入的文本数据对应到一个或多个预定的意图上，并标记文本数据所属意图类别的功能。

测试方法：按照表2已定义场景/业务文本数据的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.2 命名实体识别

测试内容：检查被测系统是否提供在文本中找出并准确标注命名实体的功能。

测试方法：按照表1命名实体文本的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.3 敏感信息辨别

测试内容：检查被测系统是否提供根据上下文对输入文本中的敏感内容进行分辨的功能。

注：敏感内容包括涉及黄色、暴力、恐怖和国家安全等信息的内容。

测试方法：按照表1敏感信息文本的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.4 语义拒识

测试内容：检查被测系统是否提供对无法处理或不应当处理的内容输入进行分辨和拒识的功能。

注：无法处理的内容包括被测系统不支持的业务内容；不应当处理的内容包括完全无意义的内容。

测试方法：按照表2未定义场景/业务文本数据的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.5 信息检索

测试内容：检查被测系统是否提供信息检索的功能，包括但不限于以下具体功能。

a) 个性化词典检索：联系人列表、歌曲列表和兴趣点（POI）等。

b) 第三方信源检索：天气、航班、酒店和股票等。

c) 自定义知识库检索。

测试方法：按照表2已定义场景/业务文本数据的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.6 文本相似度

测试内容：检查被测系统是否提供根据输入的文本数据，计算其与已有文本的语义相似度的功能，包括但不限于以下具体功能。

a) 语义不变，对句子进行替换同义词等简单的变化。

b) 语义不变，对句子结构进行变化。

c) 语义不变，改写句子的结构和用词。

d) 用词和结构非常相似，但是语义不同。

测试方法：按照表2已定义场景/业务文本数据的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.7 文本修改

测试内容：检查被测系统是否提供对对话中的前一句文本进行修改的功能。

测试方法：按照表2已定义场景/业务文本数据的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.8 语义修正

测试内容：检查被测系统是否提供对语义理解错误的结果进行自动校正的功能。

注：语义理解错误包括句法错误、中文分词错误、指代消歧错误等。

测试方法：按照表2已定义场景/业务文本数据的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.9 自然语言生成

测试内容：检查被测系统是否提供根据语义理解结果生成自然语言文本，符合说话人意图、满足语音交互响应的功能。

注：自然语言文本内容包括：

a) 简单答复文本，

b) 根据预定义模板的答复文本，

c) 理解和符合说话人意图的答复文本，

d) 说话人意图不明确时给出的合理的引导或推荐答复文本。

测试方法：按照表2已定义场景/业务文本数据的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.10 逻辑推理

测试内容：检查被测系统是否提供对文本内容的逻辑计算和推导的功能。

示例：2020年是闰年；爸爸的妈妈叫奶奶。

测试方法：按照表2已定义场景/业务文本数据的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.11 对话引导和推荐

测试内容：检查被测系统是否提供根据说话人意图和场景需求动态生成引导提示用语的功能；检查被测系统是否提供根据当前用户意图和历史用户画像，主动向用户推荐其可能感兴趣的信息，包含但不限于以下具体内容。

a) 个性化词典。
b) 根据用户行为习惯挖掘归类的信息。
c) 已定义知识库内的知识。
d) 第三方信源信息。
e) 海量数据的检索得到的关联信息。

测试方法：按照表2的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.12 上下文相关的多轮会话

测试内容：检查被测系统是否提供上下文相关的多轮会话处理能力，包含但不限于以下具体功能。

a) 对话状态跟踪。
b) 对话策略管理。
c) 对话意图切换、跳转。
d) 历史信息继承。

测试方法：按照表2已定义场景/业务文本数据的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

7 性能测试方法

7.1 理解效果

测试内容：理解效果测试项包括精确率、召回率、拒识率、准确率、F1值、平均排序倒数和归一化折损累计增益等参数，具体参数描述如下。

a) **精确率：**检测被测系统语义理解的能力，即被测系统对有效文本实际响应正确的次数占所有文本响应正确的总次数的比率。参数值计算方法见公式（1）：

$$P_{SS}=\frac{N_{SS}}{N_S}\times 100\% \quad \text{（式1）}$$

式中：

P_{SS}——语义理解精确率；

N_{SS}——有效文本实际响应正确的次数；

N_S ——所有文本响应正确的总次数。

b) **召回率**：检测被测系统语义理解的能力，即被测系统对有效文本实际响应正确的次数占应该响应正确的总次数的比率。参数值计算方法见公式（2）：

$$R_{SS}=\frac{N_{SS}}{N_{SC}}\times 100\% \quad \text{（式2）}$$

式中：

R_{SS}——语义理解召回率；

N_{SS}——有效文本实际响应正确的次数；

N_{SC}——有效文本应该响应正确的总次数。

c) **拒识率**：检测被测系统语义拒识的能力，即被测系统对无效文本实际响应正确的次数占无效文本输入的总次数的比率。参数值计算方法见公式（3）：

$$SR=\frac{N_{SR}}{N_R}\times 100\% \quad \text{（式3）}$$

式中：

SR——语义拒识率；

N_{SR}——无效文本实际响应正确的次数；

N_R ——无效文本输入的总次数。

d) **准确率**：检测被测系统语义理解的能力，即被测系统对所有文本实际响应正确次数占所有文本响应的总次数的比率。参数值计算方法见公式（4）：

$$A_{SS}=\frac{N_{SS}+N_{SR}}{N}\times 100\% \quad \text{（式4）}$$

式中：

A_{SS}——语义理解准确率；

N_{SS}——有效文本实际响应正确的次数；

N_{SR}——无效文本实际响应正确的次数；

N ——所有文本响应的总次数。

e) **F1 值**：检测被测系统语义理解的能力，即被测系统精确率和召回率的加权调和平均值。参数值计算方法见公式（5）：

$$F1=\frac{2\times P_{SS}\times R_{SS}}{P_{SS}+R_{SS}}\times 100\% \quad \text{（式5）}$$

式中：

$F1$——被测系统F1值；

P_{SS}——语义理解精确率；

R_{SS}——语义理解召回率。

f) **平均排序倒数**：检测被测系统信息检索的能力，即正确结果在被测系统给出结果中的排序位置倒数的平均值。参数值计算方法见公式（6）：

$$MRR=\frac{1}{Q}\times\sum_{i=1}^{Q}\frac{1}{rank_i} \quad \text{（式6）}$$

式中：

MRR ——平均排序倒数；

Q ——信息检索的总次数；

i ——第i次信息检索；

$rank_i$——在第i次信息检索中正确结果出现的排序位置。

g) **归一化折损累计增益**：检测被测系统信息检索的能力，即被测系统给出结果的排序相关性评分与理想结果的排序相关性评分的比值。参数值计算方法见公式（7），公式（8）和公式（9）：

$$DCG = \sum_{j=1}^{K} \frac{rel_j}{\log_2(j+1)} \quad \text{（式7）}$$

式中：

DCG——折损累计增益；

K ——信息检索结果个数；

j ——第j个检索结果；

rel_j——第j个检索结果的相关性评分。

$$IDCG = \sum_{j=1}^{|REL_K|} \frac{rel_j}{\log_2(j+1)} \quad \text{（式8）}$$

式中：

$IDCG$ ——理想结果折损累计增益；

$|REL_K|$——信息检索结果个数按照相关性评分从大到小排序；

j ——第j个检索结果；

rel_j ——第j个检索结果的相关性评分。

$$NDCG = DCG/IDCG \quad \text{（式9）}$$

式中：

$NDCG$——归一化折损累计增益；

DCG ——折损累计增益；

$IDCG$——理想结果折损累计增益。

测试方法：理解效果测试可根据不同功能选择适用测试指标进行测试，不同功能及其适用的效果测试指标对应情况见表3。

表 3　不同功能及其适用的效果测试指标

功能	精确率	召回率	拒识率	准确率	F1 值	平均排序倒数	归一化折损累计增益	备注
意图理解	必选	必选	可选	可选	可选	/	/	单测语义抽取功能，语义信息抽取正确即为正确，不关注意图理解是否正确。
命名实体识别	必选	必选	可选	可选	可选	/	/	单测命名实体识别功能，命名实体识别正确即为正确，不关注意图理解是否正确。

表 3　不同功能及其适用的效果测试指标（续）

功能	精确率	召回率	拒识率	准确率	F1 值	平均排序倒数	归一化折损累计增益	备注
敏感信息辨别	必选	必选	可选	可选	可选	/	/	
语义拒识	/	/	必选	/	/	/	/	
信息检索	可选	可选	可选	/	/	可选	必选	
文本修改	可选	必选	可选	可选	可选	/	/	
语义修正	必选	必选	可选	可选	可选	/	/	
逻辑推理	必选	必选	可选	可选	可选	/	/	

理解效果测试方法如下。

—— 测试数据：按照表2的要求制作测试数据集，对各测试数据集所有的文本内容进行人工标注，并制作成标准结果对比文件。

—— 测试工具：符合5.2的要求。

—— 测试环境：符合5.3的要求。

—— 测试执行：按照5.4的要求对被测系统进行测试。

—— 结果判定：按照表3给出的适用关系和测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、指标项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

7.2　理解效率

测试内容：理解效率测试项包括平均响应时间、响应时间分布和识别吞吐量等参数，具体参数描述如下。

a) **语义理解平均响应时间：**检测被测系统的语义理解响应速度，即在输入一条文本后，被测系统给出该条语义理解结果的时间。参数值计算方法见公式（10）：

$$T_{ssu}^{avg}=\sum_{i=1}^{N}\left(t_u^i-t_e^i\right)/N \quad \text{（式10）}$$

式中：

T^{avg}_{ssu}——语义理解平均响应时间；

t^i_u ——得到第i条文本语义理解结果的时间；

t^i_e ——输入第i条文本结束的时间；

N ——输入文本总条数。

b) **响应时间分布：**检测被测系统的语义理解响应速度，即在输入一条文本后，被测系统给出该条语义理解结果的时间为响应时间，统计测试数据集上所有的响应时间分布及其占比情况。

c) **语义理解吞吐率：**检测被测系统的语义理解整体效率，即被测系统在单位响应时间内语义理解的时间长度。参数值计算方法见公式（11）：

$$T_{suP}=\sum\nolimits_W T_S\Big/\sum\nolimits_W T_{ssu} \quad \text{（式11）}$$

式中：

T_{suP} ——语义理解吞吐率；

W ——测试集；

T_S ——测试集上文本大小；

T_{ssu} ——测试集上响应时间时长。

测试方法：理解效率测试方法如下。

—— 测试数据：按照表2的要求制作测试数据集。

—— 测试工具：符合5.2的要求。

—— 测试环境：符合5.3的要求。

—— 测试执行：按照5.4的要求对被测系统进行测试。

—— 结果判定：按照测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、指标项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

7.3 系统稳定性

测试内容：系统稳定性测试项包括稳定运行和资源使用等参数。

a) **稳定运行：**检测在给定的软硬件配置和系统并发路数的条件下，被测系统运行 4.2.1～4.3.12 的各项功能，未出现崩溃、假死或功能异常，能持续正常运行的能力。

注：给定的软硬件配置和系统并发路数需满足被测系统正常运行的能力。

b) **资源使用率：**检测在给定软硬件配置和系统并发路数的条件下，被测系统运行 4.2.1～4.3.12 的各项功能，系统物理内存、虚拟内存、CPU、GPU、句柄等各项资源使用率持续平稳的能力。

注：给定的软硬件配置和系统并发路数需满足被测系统正常运行的能力。

测试方法：系统稳定性测试方法如下。

—— 测试数据：按照测试项要求准备测试数据集，并明确软硬件配置和系统并发路数。

—— 测试工具：符合5.2的要求。

—— 测试环境：符合5.3的要求。

—— 测试执行：按照5.4的要求对被测系统进行测试，在线场景下持续7天、离线场景下持续3天连续不间断向被测系统循环输入测试文本，连续监测系统运行情况和物理内存、虚拟内存、CPU、GPU、句柄等资源使用率变化情况。

—— 结果判定：按照测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、软硬件配置、系统并发路数和指标项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

附 录 A
（资料性）
主观体验测试

A.1 概述

根据不同场景/业务需求和不同使用者感受，语义理解相关功能的效果会产生差异，本附录给出了人工主观体验测试的测试项和测试方法。

A.2 测试项

A.2.1 平均对话轮数

检测被测系统理解说话人意图所需要的平均对话轮数。计算方法见公式（A.1）：

$$R_{AVG} = \sum_{n=1}^{N_D} R_D / N_D \quad \text{（式A.1）}$$

式中：

R_{AVG}——平均对话轮数；

N_D ——请求对话总次数；

R_D ——每一次对话的交互轮数。

A.2.2 任务完成率

检测被测系统对话过程中任务完成率。计算方法见公式（A.2）：

$$R_{REACH} = \sum_{n=1}^{N_T} R_T / N_T \quad \text{（式A.2）}$$

式中：

R_{REACH}——任务完成率；

N_T ——请求对话总次数；

R_T ——每一次对话中任务达成数，一次对话中可以有多个任务。

A.2.3 满意度

每个测试人员完成体验后，从任务完成情况、响应速度等方面综合考虑，对语义理解（子）系统给出满意度等级。一般分为：非常满意、满意、一般、不满意，很差。

A.3 测试方法

A.3.1 测试数据集

在测试开始前，根据场景/业务需求，以随机产生的方式预定义说话人意图，并通过人工编写或采集的方式制作测试数据集。

A.3.2 测试工具

使用可编程测试工具和日志分析统计工具进行测试。

A.3.3 测试环境

根据被测系统的功能和性能要求以及使用场景进行软硬件环境配置。

A.3.4 测试执行

让不少于20名，不同性别、年龄段、学历背景的测试人员根据测试数据集与智能语音交互系统进行对话，记录一次对话中的交互轮数、一次对话中任务是否达成、体验完成后对整个体验过程的满意度。同时，对交互过程中的日志进行统计分析。

A.3.5 结果判定

测试结果：任务完成率越高越好；任务完成率相同，平均对话轮数越少越好；任务完成率和平均对话轮数均相同，满意度越高越好。

参 考 文 献

[1] GB/T 5271.29—2006 信息技术 词汇 第29部分：人工智能 语音识别与合成