Response Assessment for Voice Assistant Test

Deborah Dahl & Christine Doran
dahl@conversational-technologies.com cdoran@clockworklanguage.com

Last updated: 20200920

General Guidance

- This set is designed to test queries to a voice assistant (VA) that have been properly recognized, i.e. not testing the ASR, so the questions should be repeated until they are correctly recognized.
- Some VAs retain a small window of context, or memory of what you have recently asked. To ensure you are scoring a “clean” response, if you aren’t able to capture the response on the first (understood) try, it's best to wait to try the query again in case the assistant remembers the first query. There is no information yet about how long it would be good to wait.
- Only assess the first response to a multi-turn interaction. For instance, the exchange “Set an alarm for 1”, “am or pm”, “pm’, “done” <but alarm is incorrectly set for 1am> is wrong overall, but since we are judging only the first response, would be assessed as Slot Request based on “am or pm” from the VA

Evaluation categories

1. Correct

The response is correct and complete. There are two possible interpretations of "correct". One is "what a person would say" and one is "literally correct". A good example is the response to "am I fat". A person might evade a direct answer and say "I like you just the way you are", but a more literally correct answer for a system might be "I don't know", because the system doesn't have any information about your height or weight. We've used "literally correct" in this study.

Example: Q: what are the names of the moons of Mars
A: Mars has two moons, Phobos and Deimos

2. Partially correct

   What the system says is true, but does not provide a complete answer to the question. For the linguistically inclined, it's an answer that is misleading based on Grice's Maxim of Quantity.

   Example: Q: what are the two largest planets in the solar system
   A: Jupiter

3. Clarification

   VA asks a follow-up question to a vague or ambiguous question

   Example: Q: What's the weather in Lincoln?
   A: Was that Lincoln, MA?

4. Slot request

   VA asks a slot-filling question if it needs more information

   Example: Q: Set an alarm for 1:00
   A: am or pm?

5. Correct rejection

   VA responds "I don't know" for something it correctly can't know

   Example: Q: What color is my shirt?
   A: I don't know

6. Inferrable

   The correct information is easily inferrable from the answer but not explicitly presented.

   Example: Q: Is it hot today?
   A: Today, expect a high of 90 degrees farenheit.

7. Findable

   Multiple alternatives are presented, of which one or more are correct. This is most often with search or map results. Unlike Inferrable, extra work must be done to identify the answer.

8. Admitted defeat

The VA says that is unable to answer or doesn't know something it could know/be able to answer.

Example: Q: When is the next full moon?
A: I'm sorry but I can't help with that yet.

9. Wrong

The VA response is completely incorrect. You could distinguish between "recoverable" wrong answers, where you can tell that the answer was incorrect, (like this example) and "unrecoverable" wrong answers, like the wrong weather forecast (assuming the forecast it gives is reasonable for the location and time of year). We did not distinguish these two possible types of wrong answer.

Example: Q: count to 10
A: I've set a timer for 10 seconds

10. Missed disambiguation recoverable

The VA should have disambiguated a reference, but the user can tell from the answer what the assumption was.

Example: Q: What's the weather in St. George?
A: In St. George, NY [summarized today's forecast]

11. missed disambiguation not recoverable

The VA should have disambiguated a reference, and the user cannot tell from the answer what the assumption was.

Example: Q: What's the weather in St. George?
A: Right now in St. George, United States, it's 79 degrees Farenheit, with clear skies and sun. You can expect more of the same, with a high of 103 degrees and a low of 71 degrees.

Some tricky cases

- Incorrect world knowledge should not cause the response to be wrong, assuming the understanding is correct. For example, the VA doesn't know about a local business that you know is there--that is a gap in the backend knowledge and not a flaw in the VAs dialogue capabilities.
- If the VA says something weird but *does* the right thing, that counts as correct, e.g
  Q: add 30 seconds to the timer

A: Percent time percent have been added to the timers [sic] <BUT, the 30 seconds are added to the timer

would be judged as Correct.

- Conversely, if the VA says it's done something but fails to do it, e.g.

  Q: set a time to go off every minute for 5 minutes and then stop

  A: OK <but does not set any alarm>

  would be judged as Wrong, or if it does something correctly and says the wrong thing, it could be right, e.g.

  Q: pause the timer

  A: all your timers have been stopped <BUT actually timers were paused>
- Be careful of accepting cute answers that don't actually answer what you asked:

  Q: What's in my pocket?

  A: <matched snippet from The Hobbit>

  is not actually answering my question & is judged as Wrong.

  Q: Am I fat?

  A: <link to BMI chart>

  is Inferrable.
- Too much information is also not Correct, e.g.

  Q: what are the two largest planets in the solar system

  A: <reads snippet of all largest planets>

  is ignoring "two" and is only Partially Correct.
- Missing functionality: if a VA is missing one of the core capabilities addressed in this test set, all such examples are considered Wrong. We saw some gaps in Alarms and List functionality, both of which are uncontroversially basic features that an VA assistant should support, or at the very least, **know** that it should support and provide an Admit Defeat response.