Testing Bots with Bots

Conversational CX Gives Rise to a New Breed of Tester

Mike Monegan
VP Product Management
March 11, 2019
54% of US online consumers expect interactions with customer service chatbots to negatively affect their quality of life.

-- Forrester Infographic, Jan. 2019
1. Sources of complexity are intertwined in customer service chatbots & voicebots

2. Domain expertise represented in a few “Gold-standard Flows” that are the tip of the iceberg

3. Business and Technology silos impede the process of training bots well
Why use Bots to test Bots?

1. Massive scale of test coverage problem space suggests use of similar techniques that bots use
   • Sizable number of intents bots are confronted with
   • Broad range of human inputs (both spoken and textual)
   • Stakes are high with users already weary of #chatbotfails

2. Early results suggest ML & NLP drives testing productivity
   • Humans produce a few conversational variants, but machines produce interaction perturbations at scale
   • Automated cross-validation of training data & test data
   • Intent confusability testing uncovers issues that would only have been found by customers

3. Data-driven testing efficiency is learnable with supervision
   • Combinatorics of entity values makes even automated testing prohibitively slow
Many Humans In The Loop

• Both Bot Users and Human Agents must be connected efficiently
  • Escalated conversations must route to the correct agent at the correct time
  • Correct, complete context transfer is critical (many bots manage context poorly)
  • Poor handoffs can be flagged in real time by receiving agent

• Training of bots by domain experts is an on-going process

• HITL-oriented portal facilitates supervised learning
  • Intuitive, interactive design draws in and engages experts to train the models
  • Experts quickly judge conversations with accept/reject & pass/fail responses
  • Human inputs are retained to train future syntheses
Machine-Generating Bot Conversations

1. "Gold" Training Data
2. Generate Conversation Perturbations
3. New Test Case Executions
4. Process HITL Review
   - Valid test case
   - Bad bot response
5. Expert Guidance
6. Expanded Training Data
7. Expanded Test Cases
Finding the Holes in AI Training Data

• NLU & Cognitive Services available to all, but results are often quite fragile
  • Slight differences in input cause confidence drop or wrong intents
  • Results can be particularly bad if training data not clean – “Hi Micheal”
• Chatbots must handle the full range of other human foibles in textual input
  • Capitalization issues – “can you cancel my flight to ANchorage?”
  • Bad or missing punctuation – “thanks for that also need my balance”
  • Typos & abbreviations – “want to to open a savings acct”
  • Unknown words – “that’s hella bad”
• Productive to deeply test combinations of prefixes (e.g., “I need to”) & suffixes (e.g., “please”) as well as different storytelling preambles
A New Breed of Tester Emerges

• Intents, entities, workflows, user personas, and data-driven variations all point to a need for an assistant to manage complexity

• A well-equipped tester now supervises a very sophisticated testing platform and *judges* if the bot is performing appropriately

• “Tester as Teacher” role seriously raises the required expertise

• Agile iteration favors the tester with analytical & trend spotting skills

• Trend Reports on quality metrics have become very important

• Success metrics are not the same for every bot
  • Containment
  • Correctly routed
  • Intent Resolution
  • Customer Satisfaction (e.g., NPS)