Towards Turing Test 2.0:

A Product-Driven Approach to Measuring Conversational AI Using Wisdom of the Crowd

Ilya Eckstein, Robin Labs

ilya@robinlabs.com

We propose a new approach to the unsolved problem of benchmarking conversational AI agents. Our solution leverages the popular consumer-focused voice assistant Robin as a plug-your-own-bot platform for cross-sectional benchmarking using a crowd of motivated live users, with several metrics based on implicit and explicit user experience signals. We believe that this open platform, coupled with a statistically sound methodology, can help close the feedback loop for conversational NLU researchers, thereby significantly accelerating progress in the field.

Problem

Conversational intelligence is one of the cornerstone problems in AI. Indeed, the classic Turing Test practically equates AI with a machine's ability to conduct a human-like conversation, and the recently reinvigorated interest in "deep" AI has also led to new research directions in natural dialogue agents [1-7]. Further beyond the academia, the ongoing shift to conversational interfaces - both voice and chat based - has already reached the mainstream, resonating with the public (see Figure 1) while triggering a strong demand for more capable language understanding technologies. Yet, this demand remains largely unmet. Why?

Beyond the fundamental complexity of the human conversation, we hypothesize that the challenge is aggravated by the lack of adequate methods to measure performance of conversational models, hampering progress in the entire field.

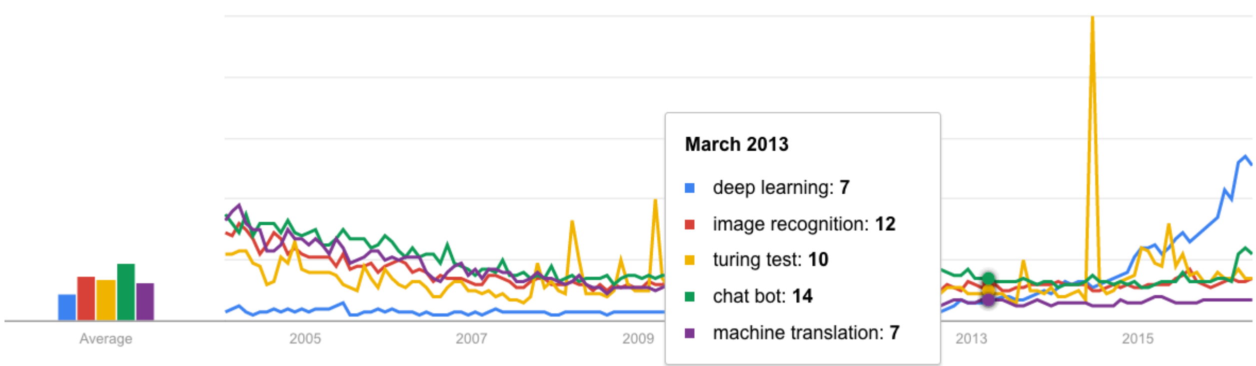

Figure 1 Current AI zeitgeist as seen through the lens of Google Trends. Topics such as "chat bot" and "Turing test" are among the top search terms in relation to AI in the last decade. Presently, "chat bot" is second only to "deep learning" of all the select topics.

Desiderata

First, let us outline our goals and key criteria for an agent evaluation approach:

★ We seek to offer an open evaluation platform to the community, a kind of chatbot playground where researchers could easily plug in and test their agents.
★ The platform should enable transparent community access to evaluation results.
★ Furthermore, all submitted chatbots should be accessible for direct interaction.
★ We prioritize real-world interactions with live users over synthetic and static tests.
★ The platform should be conducive to iterative experimentation, with all the conversation logs made openly available to the community.
★ Finally, we focus on fundamental methods and models rather than solutions tailored to specific tasks. Thus, while our toolkit can include narrow tasks, it should be capable of evaluating a broad variety of agents whose only common denominator is stepwise operation with text input and text output.

With these in mind, let us proceed.

Proposed Solution

Recognizing the futility of the quest to devise a set of gold standard input/output pairs for natural dialogue, we opt for a radically different approach. Instead of using a fixed testing dataset (or a fixed focus group), we propose to leverage an ever-active crowd of live users who are motivated to interact with chatbots. Given access to a sufficiently large audience of motivated online testers, we can then A/B test[1] different agents by tracking a number of metrics such as user engagement, retention, satisfaction, etc. In other words, instead of (or in addition to) trying to programmatically quantify "goodness" of artificial dialogue, we let live subjects (randomly drawn from the same audience pool) "vote with their feet". The end result is an "apples-to-apples" comparison between chatbots.

In truth, such cross-sectional multivariate testing is not new in principle and is commonly used in commercial product development. We are merely borrowing this approach for an open research framework. However, the key challenge is to a acquire reliably sufficient number of active, ever-motivated users - and therein lies our main contribution. Thankfully, Robin Labs is fortunate to have the asset of Robin - and access to its users. Robin is a popular voice assistant application on Android, with some 2 million downloads and used daily by thousands[2]. While its stated purpose is task-driven assistance, nearly 40% of the interactions boil down to miscellaneous chats with the bot. The motivation that drives users towards such interactions can (and should) be leveraged for research - this is the key point of our proposal.

We will now briefly review existing methods and their limitations, before delving into the details of our proposed methodology and metrics.

Prior Art

Probably the first known measure of machine intelligence, the classic Turing Test, focuses squarely on a machine's conversational abilities. Alas, the Turing Test is now widely seen as a flawed metric as, by design, it

---

[1] We use the term "A/B testing" loosely, which in general implies multivariate testing once there are more than two agents to be compared.
[2] Specific numbers will be made available upon request.

incentivizes the Artificial Conversational Agent (a.k.a chatbot) to trick the user. Many other methods have been proposed to measure an agent's ability to track dialogue states and/or to reach predefined goals [9-11], however they don't generalize to other domains or to measure the quality of broader conversational models such as the recent end-to-end methods, e.g. [1-3]. Those rely either on very few human scores [1], crowdsourcing [2] or machine translation metrics like BLEU [3] to judge only the quality of the generated responses and shown [12] to be less than adequate for chatbot benchmarking. As a result, Su et al. [8], among others, choose to evaluate their chatbots using focus groups, a method that is unfortunately not amenable to 3rd party comparison, thus not as productive for the research community as a whole.

Methodology and Metrics

Among Robin users, we broadly classify their motivation as task-driven vs. communication-driven. Users in the latter category mainly seeks to chat with the bot for the sake of chatting[3]. They are flexible about the topics and the conversations are not tied to a certain well-defined task, making it possible to switch the agents covertly without a significant degradation in user experience. Thus, our method implies randomly drawing a subset of those users and matching them with 3rd party agents that have been submitted by researchers for benchmarking. This strategy ensures that every bot will get a sample of users to interact with. But how can we evaluate those interactions? Based on our product experience, we propose the following simple metrics:

E-Metric: User Engagement One very simple user engagement metric is average conversation length, measured as the number of dialogue turns. However, even in open-domain systems, conversation length is hardly a good measure of communication efficiency as dialogues can become excessively long simply due to e.g. mounting confusion. We therefore opt for the so-called *amended conversation length* measure where we seek to penalize confusion, repetitions and other artifacts by detecting specific patterns of disapproval in users' reactions to bot responses, such as:

a. Users explicitly telling the bot about their intent being misinterpreted (e.g., *"you didn't get it"*, *"i didn't ask for ..."*, *"you are stupid"*, etc.)
b. Users repeating their request twice or more in a row, in identical or similar terms
c. Unclassified negative sentiment towards the bot, carrying a relatively lighter penalty.

Examples 1 and 2 in Appendix A illustrate this conversation scoring approach: by default, each bot response gets a score of 1, except for those tagged as *bot_confused*, resulting in a -2 score, or simply as *negative* (-1 score), etc. These examples show how relatively long exchanges can still receive negative scores.
Similarly, we can discount the score of trivial "dialogue glue" responses employed by a chatbot, such as confirmation questions, repetitions, etc. A bot that merely keeps repeating its (or the user's) words should receive a zero engagement score.

S-Metric: Implicit User Sentiment The above E-Metric already leverages sentiment analysis. Still, there is value in considering user sentiment towards the bot in an isolated fashion.

[3] What drives such user behavior is a fascinating discussion beyond the scope of this text.

X-Metric: Explicit User Rating Since we have the luxury to control all aspects of Robin, we can also go beyond implicit sentiment and ask users explicitly to rate their experience: from select conversations to individual bot responses. We also plan to experiment with Facebook-style reaction buttons (see Figure 2) to extract more specific *k*-dimensional[4] feedback from users. While, in theory, we could also engage 3rd party reviewers (e.g., via Mechanical Turk) to reduce bias in conversation assessments, such reviewers reportedly [8] lack the first-hand motivation that actual users have, making them less effective.

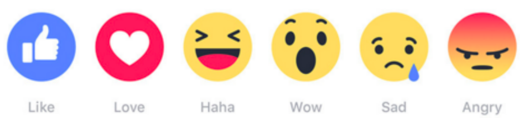

Figure 2: Facebook reaction buttons. We plan to experiment with a similar concept to solicit *k*-dimensional feedback from users

R-Metric: User Retention A common metric in tracking usage, retention is simply the percentage of new users who stick with the product after a day/week/month/etc. However, this metric implies a situation where users have an actual need that the given product is meant to satisfy. This may not be the case with experimental agents in general and open-domain chatbots in particular - a weakness we address in the next section.

In general, rather than trying to devise a single perfect metric (in reality, likely no such metric exists), our goal is to create a flexible benchmarking platform that allows researchers to develops and plug in new metrics. What we propose here is merely the initial toolset.

Open-Domain vs Domain/Task-Specific Agents

As stated in the Desiderata, our goal is to enable evaluation of fundamental methods rather than narrow, task-specific agents. The reason is that understanding task-specific commands within a narrow domain (e.g., calendar management or airfare booking), can typically be achieved with tailored solutions do not generalize well. However, our ambitiously broad goal has two shortcomings. First, the notion of a successful dialogue exchange is not well defined - although this issue can be remediated by tracking usage metrics as explained above. Perhaps more importantly, users may lack a long-term motivation to interact with agents without a clearly stated purpose. To avoid these pitfalls, we relax the original constraint as follows:

1. We identify three main categories of agents:
    a. Task-specific agents, with relatively short conversations that have a well-defined goal and therefore a clear definition of success/failure. Examples include restaurant search and flight booking;
    b. Domain-specific agents,, but without necessarily focused on a single narrow task. One example is a content recommendation agent. Consider the fact that many people read movie reviews for education or entertainment, without having an immediate intent to watch those movies;
    c. Open-domain agents.
2. In each of the above categories, we plan to specify one or two specific agent subcategories. Each submitted agent will have to match a subcategory.
3. We allow submissions of domain/task specific agents as long as the underlying algorithm is a general one, i.e., not engineered specifically for the given task. One way of enforcing this constraint requiring for

[4] We plan to experiment with emoji buttons matching four common reactions: *Like, Dislike, Funny* and *What the [...] was that?*

each agent to be submitted to at least two categories - with the same code/binaries but potentially different data models. This process is meant to facilitate generalization abilities of each agent.

Incidentally, we believe that the domain-specific agents currently represent a “sweet spot” for research, as they are somewhat constrained while still being conducive to rich conversations, particularly when the agent serves a long-term purpose rather than a one-off function. Again, one qualified example is personalized movie recommendation agents. Indeed, in the conversational medium, personalization typically involves fairly long, loosely constrained user-agent conversations that nevertheless have a clear purpose. Here is a hypothetical example of a such dialogue:

User: Can you suggest a good comedy to watch?
Bot: Sure, but first I need to learn about your preferences. Can you tell me more about yourself?
User: Ok. I am 38 years old, speak English, Russian and some French. I love theatre, dance, and art house cinema.
Bot: Art house, as in ...?
User: As in French New Wave, for instance.
Bot: Something like Godard?
User: Early Godard, definitely! Breathless is one of my favorites!
Bot: Oh yes, Belmondo and Jean Seberg, right?
User: Exactly! I never get tired of it! Also, Fellini
Bot: What do you like about Fellini?
User: I don't know... everything! The look, the feel, the rhythm, the actors. It's just unique
Bot: Interesting... So far I got Godard, Fellini, Belmondo, Seberg...
User: Yup
Bot: How about something contemporary?
User: Wes Anderson is good
Bot: Hmm... have you seen his “The great Budapest Hotel”? It's got a great cast!
User: Wasn't it Oscar-nominated?
Bot: Yes it was, didn't win Best Picture though. But still, it's widely acclaimed!
...

In our product experiments with Robin, we have learned that content recommendations are a genuine use case: many users are willing to accept recommendations to from a bot and are strongly motivated to invest effort in personalization[5].
Our interest in conversational recommendations for chatbot benchmarking is not unique: Dodge et al. [13] recently proposed a suite of movie-themed tasks that is essentially a conversational recommendation system. However, instead of using their fixed dataset of of Q&A pairs that is tied to a very specific choice of words and recommendation policies, we prefer to remain within our suggested evaluation framework. Also, to decouple dialogue assessment from the quality of underlying recommendations, we plan to offer a personalized movie recommendations API, trained on MovieLens [14] data, as part of our toolkit.

In summary, we plan to introduce vertical agent categories in the platform, starting with movie recommendations, explicitly promoting them to Robin users and allowing researchers to target specific categories. More categories will likely be added in the future.

[5] More details can be shared upon request.

Voice vs Text Interface

In our experience, mainstream users are more likely to have longer conversations with a voice interface rather than with a text-based chatbot. The caveat is that the interface needs to be reliable and well-designed, i.e., have a pleasant synthetic voice and speech recognition must not be a barrier. Fortunately, that groundwork is already established in Robin, and Android's built-in speech recognition now works extremely well for native English speakers. We therefore believe that Robin's voice-first interface provides a low-friction starting point for testers.

Towards Reinforcement Learning of Dialogue Models

Supervised learning approaches, partucularly those based on neural networks, often require large amounts of annotated training data, a limitation rendering them close to impractical in many cases. As a result, reinforcement (and, more generally, reward-based) learning is emerging as a viable alternative to optimization-based conversational modeling [8, 15-17], but it relies on the ability to collect obtain reward in response to agent actions, typically in the form of user feedback. Thus, we anticipate that our framework's ability to collect such reward from users can be particularly instrumental in reinforcement/reward-based learning of dialogue.

Summary

We propose a radically new approach to benchmarking conversational AI agents by allowing live cross-sectional testing against real users of Robin - a popular consumer product. We also propose an initial (but by no means exclusive) set of metrics based on tracking user behavior and leveraging implicit and explicit user feedback. Researchers will be able to submit and assess agents in two categories: open-domain chat and personalized movie recommendations, with more categories to be added in the future. Our long term goal is to leverage Robin as a platform for experimental chatbots to accelerate progress in the field.

Appendix A: Examples of (real) user vs. bot exchanges scored with the E-Metric

Example 1

User utterance	**Bot response**	**User Reaction Labels**	**Response score**
my name is royal not 35 to	*Nice to meet you Royal Not 35 To! Did I get your name right?*		1
no you didn't get my name right and i cannot hear you	*Whatever you say!*	bot_confused	-2
okay you need my name let me tell it to you	*Sure, as you say!*		1
you still didn't get my name right	*${username}! That's a lovely name!*	bot_confused	-2
that's not my name	*Please write to my parents. They are working to make me better.*	bot_confused	-2
you need to change my name is robyn	*Nice to meet you Robin! Did I get your name right?*		1
no you didn't	*Cancelling it*	negative	-1
your welcome and i can't hear you	*Ok, raising volume*		1
do you know my name now	*You are ${username}, we make a great team!*		1
***yes** we do **thank you** for getting my **damn** name right*	*Right away!*	approval negative	2-1=1

Total Score	**-1**

Example 2

User utterance	**Bot response**	**User Reaction Labels**	**Response score**
june 13 calendar	*Showing your calendar meetings on Saturday June 13*		1
what the hell are you talking about	*I think I will need to talk to your mom*	bot_confused	-2
your smart ass you know that	*You really need to go out more!*		1
no you need to go out more	*Ok, I am not doing it*		1
why the hell not	*Oooh!*	negative	-1

you suck you know that right	*Did I tell you my middle name was Hoover? Seriously, can we try again?*	negative	-1
yo mama	*Hi ${username}! What would you like to tell me?*		1
do not want to know damn you	*I am sorry, I am learning as I go. Could you rephrase please?*	bot_confused	-2
i do not want to listen to the music	*Whatever you say!*	bot_confused	-2
whatever whatever	*Ok, I'll stop*		1
you need to listen carefully	*I'm here for you*		1
you're not listening to me	*Sure, I am listening to you.*	bot_confused	-2
what's your name	*I am Robin. And what about you?*		1
big dick johnson	*You really need to go out more!*		1
***i said** why do you say i need to go out more*	*Before you find your handsome prince, you have to kiss lots of frogs.*	bot_confused	-2
that didn't make sense	*I am sorry if I don't always understand you.*	bot_confused	-2
i'm going to go get in the shower right now so go piss off	*What are we, in the third grade?*	negative	-1
4th grade actually	*My point exactly!*		1
***your point was wrong** you said 3rd grade*	*Ok, I am not doing it*	bot_confused	-2
you always say that when you have nothing to say	*When you have nothing to say*		1
no when you	*Ok, I am not doing it*		1
so you think	*I am pondering it in my spare time*		1
you should you have plenty of it	*Showing your calendar meetings*		1
ew	*Let's find an answer to ew*		1
*it's a statement **stupid***	*I am on it!*	negative	-1
right	*Right away!*		1
right	*Sure, as you say!*	bot_confused[6]	-2

Total Score	**-5**

[6] Repetition of same/similar user command is seen by this metric as a sign of the bot being confused.

Bibliography

[1] Vinyals, Oriol, and Quoc Le. "A neural conversational model." *arXiv preprint arXiv:1506.05869* (2015).

[2] Jafarpour, Sina, Burges, Christopher JC, and Ritter, Alan. Filter, rank, and transfer the knowledge: Learning to chat. *Advances in Ranking*, 10, 2010.

[3] Sordoni, Alessandro, et al. "A neural network approach to context-sensitive generation of conversational responses." *arXiv preprint arXiv:1506.06714*(2015).

[4] Henderson, Matthew. "Machine learning for dialog state tracking: A review."*Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing*. 2015.

[5] Weston, Jason, et al. "Towards AI-complete question answering: A set of prerequisite toy tasks." *arXiv preprint arXiv:1502.05698* (2015).

[6] Kumar, Ankit, et al. "Ask me anything: Dynamic memory networks for natural language processing." *arXiv preprint arXiv:1506.07285* (2015).

[7] Yao, Kaisheng, Geoffrey Zweig, and Baolin Peng. "Attention with Intention for a Neural Network Conversation Model." *arXiv preprint arXiv:1510.08565*(2015).

[8] P-H. Su, D. Vandyke, M. Gasic, D. Kim, N. Mrksic, T-H. Wen and S. Young (2015). "Learning from Real Users: Rating Dialogue Success with Neural Networks for Reinforcement Learning in Spoken Dialogue Systems." Interspeech 2015, Dresden.

[9] Paek, Tim. Empirical methods for evaluating dialog systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9*, pp. 2. Association for Computational Linguistics, 2001.

[10] Griol, David, Hurtado, Llu ́ıs F, Segarra, Encarna, and Sanchis,Emilio. A statistical approach to spoken dialog systems design and evaluation. *Speech Communication*, 50(8):666–682, 2008

[11] Williams, Jason, et al. "The dialog state tracking challenge." Proceedings of the SIGDIAL 2013 Conference. 2013.

[12] Lowe, Ryan, et al. "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems." *arXiv preprint arXiv:1506.08909*(2015).

[13] Dodge, Jesse, et al. "Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems." *arXiv preprint arXiv:1511.06931* (2015).

[14] https://movielens.org/

[15] Weston, Jason. "Dialog-based Language Learning." *arXiv preprint arXiv:1604.06045* (2016).

[16] Su, Pei-Hao, et al. "On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems." *arXiv preprint arXiv:1605.07669* (2016).

[17] Paek, Tim. "Reinforcement learning for spoken dialogue systems: Comparing strengths and weaknesses for practical deployment." *Proc. Dialog-on-Dialog Workshop, Interspeech*. 2006.