Received: 2 August 2021 | Accepted: 23 September 2021

DOI: 10.1002/aaai.12030

RESEARCH ARTICLE

The success of Conversational AI and the AI evaluation challenge it reveals

Ian Beaver

Verint Systems Inc, Melville, New York, USA

Correspondence
Ian Beaver, Verint Systems Inc,
175 Broadhollow Rd, Ste 100, Melville,
NY 11747, USA.
Email: ian.beaver@verint.com

Abstract

Research interest in Conversational artificial intelligence (ConvAI) has experienced a massive growth over the last few years and several recent advancements have enabled systems to produce rich and varied turns in conversations similar to humans. However, this apparent creativity is also creating a real challenge in the objective evaluation of such systems as authors are becoming reliant on crowd worker opinions as the primary measurement of success and, so far, few papers are reporting all that is necessary for others to compare against in their own crowd experiments. This challenge is not unique to ConvAI, but demonstrates as AI systems mature in more "human" tasks that involve creativity and variation, evaluation strategies need to mature with them.

Conversational artificial intelligence (AI), or ConvAI as it has been abbreviated, is a subfield of AI where the goal is to build an autonomous agent that is capable of maintaining natural discourse with a human over some interface such as text or speech. The purpose may be to help humans perform tasks as a virtual/digital assistant, provide a natural language interface to another system as in information retrieval or navigation systems, or simply to converse like one would with an open domain chatbot. Such agents were traditionally called dialog systems, a term coined in the 1960s to describe early natural language interfaces (Raphael, 1964; Suppes, 1966), but around 2007 the term ConAI began appearing in literature to refer to complex dialog systems that not only responded to natural language queries but also incorporated other aspects of AI such as affective computing (Methta et al., 2007; Zhang, 2008). The term has since become generally interchangeable with modern dialog systems incorporating primarily neural architectures in their construction and they are trained on large volumes of human–human conversations to enable human-like interactions.

The field of ConvAI has exploded in recent years following the widespread adoption of virtual assistants such as Siri, Alexa, and Google Assistant which have made natural language a convenient user interface to a variety of physical and digital tasks. Users no longer have to get up to turn off the lights or manually type a reply to a text message, they can just tell their virtual assistant of choice to do it for them. This widespread public exposure to ConvAI combined with many recent breakthroughs in tensor optimized hardware[i] and neural architectures has brought increasing research interest into pushing its conversational abilities as close to that of humans as possible.

As a reviewer for dialog tracks in a variety of AI conferences I have been observing with interest a disturbing yet predictable evaluation dilemma surfacing in recent submissions and resulting reviewer discussions. Before large-scale neural text generation models were available, dialog systems used response generation techniques that typically worked by selecting from prepared responses in a database or by rendering some form of templates derived from the training data. These approaches were similar enough to information retrieval or machine translation tasks that system evaluation was performed by applying existing metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or precision @ k[ii] to system responses from a ground truth testing dataset of natural language inputs and target responses.

However, with the introduction of large neural text generation models such as OpenAI's GPT family (Radford et al., 2018) and their rapid subsequent incorporation into ConvAI systems due to the richness and expression they provide (Budzianowski and Vuli, 2019), objective and reproduceable evaluation of ConvAI systems became much harder. Indeed, it is possible for a generative model to produce an acceptable response to a query in conversation with no word overlap to the target response. Think of all of the ways different people could formulate a natural language response to a query such as "Where is the nearest coffee shop?," for example, "three blocks north of here," "on the corner of Main and Stewart streets," and others.

Recent papers usually report objective metrics on test sets as before but may quickly (and justifiably) discount the inevitable poor performance as due to creativity of the system in language production compared to the ground truth responses and include crowd sourced reviews as the qualitative metric of model performance (Budzianowski and Vuli, 2019) or may actually calculate BLEU scores against the training data to demonstrate that the system is in fact producing original text and not memorizing the training data (Geerlings and Merono-Penuela, 2020). Such papers are explicitly evaluating for richness and variation in responses which is opposite to a traditional ground truth benchmark objective comparison. While this is a natural progression of the technology and such variation leads to more human-like interactions, this dependance on crowd workers for the sole determination of success can lead to difficulty for reviewers to determine if a system is indeed an improvement over the current SOTA.

For example, a paper under review may report a group of crowd workers said their system is best at a task using their own worker evaluation criteria. While an existing paper may report an altogether different group of crowd workers said their system is best at a similar task, but the evaluation task the crowd workers performed was not sufficiently described in one or the other paper or they were described but the worker instructions differed. Either way the crowd results are not directly comparable between the papers which makes the contribution more difficult to judge, not to mention for someone else to reproduce the findings or determine which approach would be superior for their needs.

In addition, these papers may not use similar demographics or sizes of crowds, may not report details about the worker selection, may not detail the evaluation task as it was presented to workers, or define the task to workers in the same way as previous papers. This aspect of reproducibility is shared with the challenges faced in labeling data for supervised machine learning (ML) tasks in general. A recent study on that topic found that of 164 papers releasing a new dataset 75% gave some information on who did the labeling, 55% specified the number of labelers, 43% described instructions given to labelers, 15% provided some labeler training details, and 0%reported how much crowdworkers were paid (Geiger et al., 2020). These results highlight a need for stronger guidelines in publishing the human contributions to ML data preparation and model evaluation in general, but evaluation of ConvAI systems is becoming reliant on subjective human review by its very success.

Unlike merely agreeing on an appropriate class label for a segment of text, evaluating successful communication is much more nuanced. Successful communication goes much deeper than just the passing of information back and forth between conversants. As Grice argued in his Cooperative Principal – exchanges in conversation are not disconnected remarks but cooperative efforts and each conversant views them with a common purpose or direction (Grice, 1975). An utterance production in the context of a specific conversation is therefore not merely a string of tokens presented with a statistical prior. It is a collaborative effort that requires give and take from both parties as they establish and maintain common ground while trying to minimize the effort they collectively put forth to do so (Clark and Brennan, 1991). This very process involves something that is difficult to objectively measure: creativity. Moving from fact-based question answering such as "What is the capital of Uzbekistan?" ConvAI system outputs are entering the realm of creativity, incorporating large-scale language models that have even learned to convey contempt of Millennial work ethic in the tone of their responses as a result of reading the Internet[iii]. While there is a strong argument that this perceived creativity is just human mimicry (Bender et al., 2021), the challenge it poses to evaluate and compare system performance in a repeatable way remains.

This article is not intended to provide rigorous guidelines, although there is progress in this area (Geiger et al., 2021; Gundersen, 2021), but instead to draw attention to the problem behind the problem. With ConvAI fast approaching the communicative abilities of humans in many tasks[iv], the creative element of conversation and the subsequent reliance on human judges to determine system performance requires researchers in this field to present their system evaluations in a reproduceable way, and program committee members need to hold authors accountable to this for the sake of scientific quality. For instance, the crowd worker evaluation task and disagreement resolution methods need be described in enough detail that theoretically another researcher could form their own similar-sized group of evaluators with similar demographics, give them the system output, describe to them the evaluation task, and come up with comparable results. If it is up to a group of humans to decide what model is better at creative and accurate production of natural language in context of conversations, the way that group decision is made and the outcome must be repeatable by others or else those of us who wish to bring to bear the current breakthroughs in science will forever be sifting through haystacks of unverifiable results looking for the needle of a true ConvAI advancement for real-world applications.

CONFLICT OF INTEREST

There is no conflict of interest.

ORCID

Ian Beaver https://orcid.org/0000-0003-0865-1214

ENDNOTES

[i] https://www.techspot.com/article/2049-what-are-tensor-cores/

[ii] precision@k = (# of results in top k that are relevant)/(# of results in top k)

[iii] https://medium.com/swlh/i-think-gpt-3-is-angry-at-me-e3f125cc2385

[iv] https://super.gluebenchmark.com/leaderboard

REFERENCES

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?." In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–23.

Budzianowski, Paweł & Ivan Vuli 2019. "Hello, It's GPT-2-How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems." In Proceedings of the 3rd Workshop on Neural Generation and Translation (pp. 15–22).

Clark, Herbert H., and Susan E. Brennan. 1991. "Grounding in Communication." In *Perspectives on Socially Shared Cognition*, 127–49. Washington, DC: American Psychological Association.

Geerlings, Carina & Albert Merono-Penuela 2020. "Interacting with GPT-2 to Generate Controlled and Believable Musical Sequences in ABC Notation." In Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA) (pp. 49–53). Association for Computational Linguistics.

Geiger, R. Stuart, Dominique Cope, Jamie Ip, Marsha Lotosh, Aayush Shah, Jenny Weng, and Rebekah Tang. 2021. "Garbage In, Garbage Out" Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data?." *Quantitative Science Studies*: 2(3): 1–32.

Geiger, R. Stuart, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang & Jenny Huang. 2020. "Garbage in, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-labeled Training Data Comes From?." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 325–36). Association for Computing Machinery, New York, NY, USA.

Grice, Herbert P. 1975. "Logic and Conversation." In *Speech Acts*, 41–58. Brill.

Gundersen Odd Erik. 2021. "The fundamental principles of reproducibility." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379 (2197): https://doi.org/10.1098/rsta.2020.0210

Lin, Chin-Yew. 2004. "ROUGE: A Package for Automatic Evaluation of Summaries." In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Methta, Manish, Stephen Dow, Michael Mateas & Blair MacIntyre 2007. "Evaluating a Conversation-Centered Interactive Drama." In Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems. IFAAMAS. New York, NY, USA: Association for Computing Machinery.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu 2002. "BLEU: A Method for Automatic Evaluation of Machine Translation." In 40th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 311–8). Philadelphia: ACL.

Radford, Alec, Karthik Narasimhan, Tim Salimans & Ilya Sutskever 2018. "Improving Language Understanding by Generative Pre-training." OpenAI.

Raphael, Bertram. 1964. SIR: A Computer Program for Semantic Information Retrieval. AI Technical Reports Boston: MIT.

Suppes, Patrick. 1966. "Plug-In Instruction." In *Readings in Classroom Management*, 271. New York: MSS Information Corporation.

Zhang, Li. 2008. "Metaphorical Affect Sensing in an Intelligent Conversational Agent." In *International Conference on Advances in Computer Entertainment Technology*. New York: Association for Computing Machinery.

AUTHOR BIOGRAPHY

Ian Beaver (PhD, University of New Mexico) has worked on topics surrounding human-computer interactions such as gesture recognition, user preference learning, and communication with multi-modal intelligent virtual assistants since 2005. He has authored nearly 40 patents within the field of human language technology and regularly serves as a PC member in many top AI and NLP conferences. Ian is Chief Scientist at Verint Systems Inc where he works to optimize human productivity in contact centers by way of automation and augmentation and improve customer self-service experiences through the application of conversational AI.

How to cite this article: Beaver, I. 2022. "The success of Conversational AI and the AI evaluation challenge it reveals." *AI Magazine* 43: 139–141. https://doi.org/10.1002/aaai.12030