ChatEval: A Tool for the Systematic Evaluation of Chatbots

João Sedoc* **Daphne Ippolito*** **Arun Kirubarajan** **Jai Thirani** **Lyle Ungar** **Chris Callison-Burch**
*Authors contributed equally
University of Pennsylvania
{joao,daphnei,kiruba,jthirani,ungar,ccb}@seas.upenn.edu

Abstract

Open-domain dialog systems are difficult to evaluate. The current best practice for analyzing and comparing these dialog systems is the use of human judgments. However, the lack of standardization in evaluation procedures, and the fact that model parameters and code are rarely published hinder systematic human evaluation experiments. We introduce a unified framework for human evaluation of chatbots that augments existing chatbot tools, and provides a web-based hub for researchers to share and compare their dialog systems. Researchers can submit their trained models to the ChatEval web interface and obtain comparisons with baselines and prior work. The evaluation code is open-source to ensure evaluation is performed in a standardized and transparent way. In addition, we introduce open-source baseline models and evaluation datasets. ChatEval can be found at https://chateval.org.

Introduction

Reproducibility and model assessment for open-domain dialog systems is challenging, as many small variations in the training setup or evaluation technique can result in large differences in perceived model performance. In addition, as the field has grown, it has become increasingly fragmented.

Papers often focus on novel methods, but insufficient attention has been paid to ensuring that datasets and evaluation remain consistent and reproducible. For example, while human evaluation of chatbot quality is extremely common, few papers publish the set of prompts used for this evaluation, and almost no papers release their learned model parameters. Because of this, papers tend to evaluate their methodological improvement against a sequence-to-sequence (Seq2Seq) baseline (Sutskever et al., 2014) rather than against each other.

Seq2Seq was first proposed for dialog generation by Vinyals and Le (2015) in a system they called the Neural Conversational Model (NCM). Due to the NCM being closed-source, nearly all the papers comparing against it have implemented their own versions, with widely varying performance. Indeed, we found no model, neither among those we trained nor those available online, that matched the performance of the original NCM, as evaluated by humans.

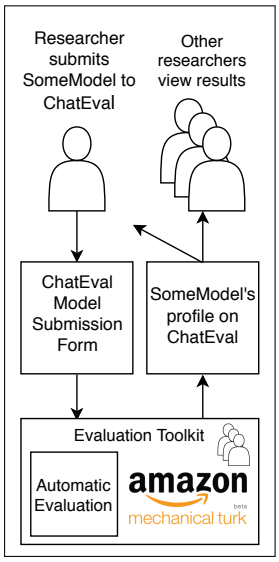

Figure 1: Flow of information in ChatEval. A researcher submits information about her model, including its responses to prompts in a standard evaluation set. Automatic evaluation as well as human evaluation are conducted, then the results are posted publicly on the ChatEval website.

Another issue is that human evaluation experiments, which are currently the gold standard for model evaluation, are equally fragmented, with almost no two papers by different authors adopting the same evaluation dataset or experimental procedure.

To address these concerns, we have built ChatEval, a scientific framework for evaluating chatbots. ChatEval consists of two main components: (1) an open-source codebase for conducting automatic and human evaluation of chatbots in a standardized way, and (2) a web portal for accessing model code, trained parameters, and evaluation results, which grows with participation. In addition, ChatEval includes newly created and cu-

Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG), pages 42–44,
Tilburg, The Netherlands, November 5 2018.

rated evaluation datasets with both human annotated and automated baselines.

Related Work

Competitions such as the Alexa Prize,[1], ConvAI[2] and WOCHAT,[3] rank submitted chatbots by having humans converse with them and then rate the quality of the conversation. However, asking for absolute assessments of quality yields less discriminative results than soliciting direct comparisons of quality. In the dataset introduced for the ConvAI2 competition, nearly all the proposed algorithms were evaluated to be within one standard deviation of each other (Zhang et al., 2018). Therefore, for our human evaluation task, we ask humans to directly compare the responses of two models given the previous utterances in the conversation.

Both Facebook and Amazon have developed evaluation systems that allow humans to converse with (and then rate) a chatbot (Venkatesh et al., 2018; Miller et al., 2017). Facebook's ParlAI [4] is the most comparable system for a unified framework for sharing, training, and evaluating chatbots; however, ChatEval is different in that it entirely focuses on the evaluation and warehousing of models. Our infrastructure relies only on output text files, and does not require any code base integration .

The ChatEval Web Interface

The ChatEval web interface consists of four primary pages. Aside from the overview page, there is a model submission form, a page for viewing the profile of any submitted model, and a page for comparing the responses of multiple models.

Model Submission When researchers submit their model for evaluation, they are also asked to submit the following: A description of model which could include link to paper or project page. The model's responses on at least one of our evaluation datasets. Researcher may also optionally submit a URL to a public code repository and a URL to download trained model parameters.

After the code and model parameters are manually checked, we use the ChatEval evaluation toolkit to launch evaluation on the submitted responses. Two-choice human evaluation experiments compare the researchers' model against baselines of their choice. New models submitted to the ChatEval system become available for future researchers to compare against. Automatic evaluation metrics are also computed. At the researchers' request, results may be embargoed prior to publication.

[1] https://developer.amazon.com/alexaprize

[2] http://convai.io/

[3] http://workshop.colips.org/wochat/

[4] https://parl.ai

Model Profile Each submitted model as well as each of our baseline models have a profile page on the ChatEval website. The profile consists of the URLs and description provided by the researcher, the responses of the model to each prompt in the evaluation set, and a visualization of the results of human and automatic evaluation.

Response Comparison To facilitate qualitative comparison of models, we offer a response comparison interface where users can see all the prompts in a particular evaluation set, and the responses generated by each model.

Evaluation Toolkit

The ChatEval evaluation toolkit is used to evaluate submitted models. It consists of an automatic evaluation and a human evaluation component.

Automatic Evaluation Automatic evaluation metrics include: The number of unique n-grams in the model's responses divided by the total number of generated tokens. Average cosine-similarity between the mean of the word embeddings of a generated response and ground-truth response (Liu et al., 2016). Sentence average BLEU-2 score (Liu et al., 2016). Response perplexity, measured using the likelihood that the model predicts the correct. response (Zhang et al., 2018). Our system is easily extensible to support other evaluation metrics.

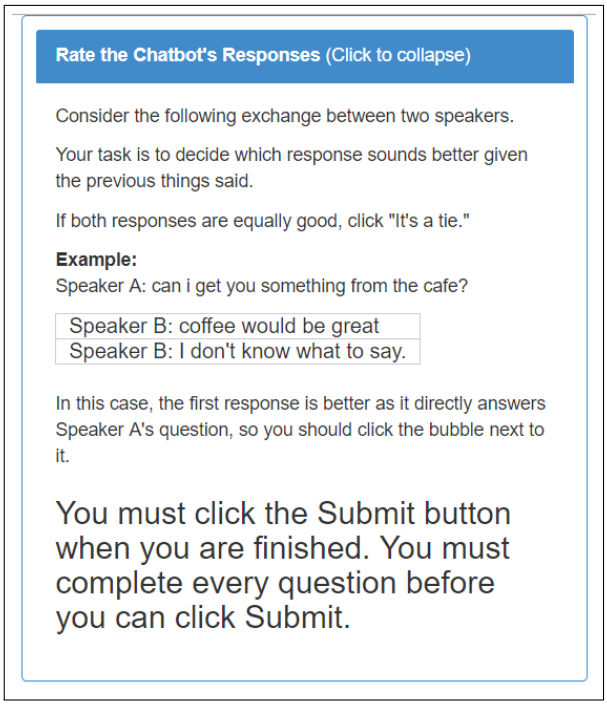

Figure 2: The instructions seen by AMT workers.

Human Evaluation A/B comparison tests consist of showing the evaluator a prompt and two possible responses from models which are being compared. The

prompt can consist of a single utterance or a series of utterances. The user picks the better response or specifies a tie. When both responses are the same, a tie is automatically recorded. The instructions seen by AMT workers are shown in Figure 2.

The evaluation prompts are split into blocks (currently defaulted to 10). Crowd workers are paid $0.01 per single evaluation. We used three evaluators per prompt, so, if there are 200 prompt/response pairs, we have 600 ratings and the net cost of the experiment is $6. On the submission form, we ask researchers to pay for the cost of the AMT experiment.

The overall inter-annotator agreement (IAA) varies depending on the vagueness of the prompt as well as the similarity of the models. Out of 18 different experiments run, we found that IAA, as measured by Cohen's weighted kappa (Cohen, 1968), varies between .2 to .54 if we include tie choices. This is similar to the findings of Yuwono et al. who also found low inter-annotator agreement. Unfortunately, there are occasionally bad workers, which we automatically remove from our results. In order to identify such workers, we examine the worker against the other annotators.

Evaluation Datasets

We propose using the dataset collected by the dialogue breakdown detection (DBDC) task (Higashinaka et al., 2017) as a standard benchmark. The DBDC dataset was created by presenting participants with a short paragraph of context and then asking them to converse with three possible chatbots: TikTok, Iris, and CIC. Participants knew that they were speaking with a chatbot, and the conversations reflect this. We randomly selected 200 human utterances from this dataset, after manually filtering out utterances which were too ambiguous or short to be easily answerable. As the DBDC dataset does not contain any human-human dialog, we collected reference human responses to each utterance.

For compatibility with prior work, we also publish random subsets of 200 query-response pairs from the test sets of Twitter and OpenSubtitles. We also make available the list of 200 prompts used as the evaluation set by Vinyals and Le (2015) in their analysis of the NCM's performance.

The datasets used for chatbot evaluation ought to reflect the goal of the chatbot. For example, even if a chatbot is trained on Twitter, it only makes sense to evaluate on Twitter if the chatbot's aim is to be skilled at responding to Tweets. With the DBDC dataset, we emphasize the goal of engaging in text-based interactions with users who know they are speaking with a chatbot. We believe that this dataset best represents the kind of conversations we would expect a user to actually have with a text-based conversational agent.

Conclusion

ChatEval is a framework for systematic evaluation of chatbots. Specifically, it is a repository of model code and parameters, evaluation sets, model comparisons, and a standard human evaluation setup. ChatEval seemlessly allows researchers to make systematic and consistent comparisons of conversational agents. We hope that future researchers–and the entire field–will benefit from ChatEval.

References

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. 2017. Overview of dialogue breakdown detection challenge 3. *Proceedings of Dialog System Technology Challenge*, 6.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, pages 2122–2132. Association for Computational Linguistics.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *EMNLP*, pages 79–84. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2018. On Evaluating and Comparing Conversational Agents. (Nips):1–10.

Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *Natural Language Dialog Systems and Intelligent Assistants*, 37:233–239.

Steven Kester Yuwono, Wu Biao, and Luis Fernando DHaro. Automated scoring of chatbot responses in conversational dialogue.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? pages 1–14.