ChatEval: A Tool for Chatbot Evaluation

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 (i.e., chatbots) 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 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 standardization and transparency. 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 significant differences in perceived model performance (Fokkens et al., 2013). While reproducibility is problematic for NLP in general, this is especially true for dialog systems due to the lack of automatic metrics. In addition, as the field has grown, it has become increasingly fragmented in human evaluation methodologies.

Papers often focus on novel methods, but insufficient attention is 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.

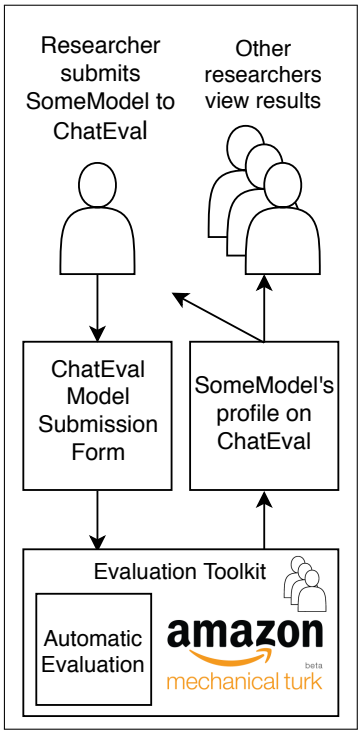

Figure 1: Flow of information in ChatEval. A researcher submits information about her model, including its responses to prompts in a standard evaluation set. Automatic evaluation as well as human evaluation are conducted, then the results are posted publicly on the ChatEval website.

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 papers compare against their own reimplementations of the model, which can vary widely in performance. Indeed, we found no model, neither

Proceedings of NAACL-HLT 2019: Demonstrations, pages 60–65
Minneapolis, Minnesota, June 2 - June 7, 2019.

among those we trained nor those available online, that matched the performance of the original NCM, as evaluated by humans.

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 curated 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 takes as input text files containing model responses and does not require any code base integration.

RankME[5] (Novikova et al., 2018) is an evaluation system for natural language generation. While RankME could be adapted for chatbot evaluation, this would require significant modification of the source code. Furthermore, RankME is only a crowdsourcing framework, which is more narrow than ChatEval. DialCrowd[6] (Lee et al., 2018) is a tool for the easy creation of human evaluation tasks for conversational agents. Finally, Kaggle[7] is another important venue for competitions, which allows for multiple test beds. However, none of these tools and websites offer a unified solution to public baselines, evaluation sets, and an integrated A/B model testing framework.

In many ways, the goal of ChatEval is similar to Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output (Federmann, 2012). Just as Appraise is integrated with WMT, ChatEval should also be used in shared tasks in dialog competitions.

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 asked to upload the model's responses on at least one of our evaluation datasets. They also submit a description of the model which could include a link to paper or project page. Researchers may also optionally include a URL to a public code repository and a URL to download trained model parameters.

After the submission is 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. Automatic evaluation metrics are also computed. If researchers opt to make their model results publicly accessible, the newly submitted model becomes available for future researchers to compare against.

Model Profile Each submitted model, as well as each of our baseline models, have a profile page on

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

[2] http://convai.io/

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

[4] https://parl.ai

[5] https://github.com/jeknov/RankME

[6] https://dialrc.org/dialcrowd.html

[7] https://www.kaggle.com

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 the 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 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: TickTock (Yu et al., 2015), Iris[8], and Conversational Intelligence Challenge[9]. 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 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.

We believe that the DBDC dataset best represents the kind of conversations we would expect a user to have with a text-based conversational agent. The datasets used for chatbot evaluation ought to reflect the goal of the chatbot. For example, it only makes sense to evaluate a model trained on Twitter using a test set derived from 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.

Overfitting One important feature of ChatEval is the ease of adding new evaluation datasets. In order to assure that researchers are not overfitting to any evaluation set, the ChatEval team will take top performing models and also apply them to other datasets. New evaluation datasets can be added upon request from the ChatEval team. We plan to add both the prompts as well as the model responses from Baheti et al. (2018) as well as Li et al. (2019). Finally, we have added the ability to interact with baseline models using FlowAI (Wubben, 2018).[10]

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:

- Lexical diversity (*distinct-n*), the number of unique n-grams in the model's responses divided by the total number of generated tokens (Li et al., 2016).
- 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).[11]

Our system is easily extensible to support other evaluation metrics.

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 model responses are exactly 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

[8] https://openi.org/solutions/iris-chatbot/

[9] http://convai.io/2017/

[10] https://flow.ai/

[11] This requires the models to be generative and publicly available code.

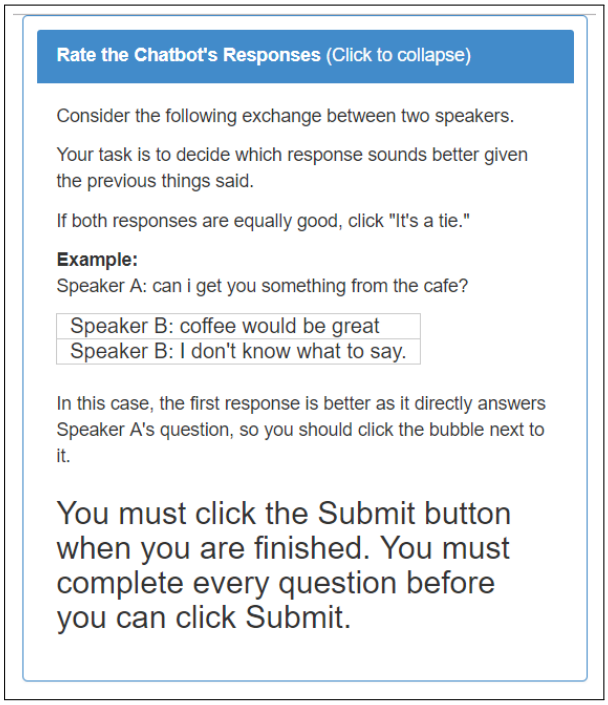

Figure 2: The instructions seen by AMT workers.

prompt/response pairs, we have 600 ratings, and the net cost of the experiment is $7.20 after fees. 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. The low IAA is similar to the findings of Yuwono et al. (2018) 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.

For analysis of relative performance between models in ChatEval, we use item response theory (IRT) to select prompts as well as test statistical significance. IRT is the basis for almost all psychometric studies (Embretson and Reise, 2013). We follow the work of Otani et al. (2016), who used head-to-head pairwise testing to compare machine translation systems. However, we further this work by also examining the discriminative power of prompts. For instance "*my name is david . what is my name ?*" from the NCM evaluation dataset has been shown to have low discriminative power, whereas, "*tell me something about your parents ?*" is useful to distinguish between the relative performance of models.

Availability of Toolkit

We expect it will be common for researchers to want to test out several of their models privately before submitting to the public ChatEval website. The ChatEval evaluation toolkit is available on Github for anyone to run.[12] We provide clear instructions for researchers to perform the human and automatic evaluation on their own with the toolkit as an alternative to using our web interface.

Availability of the Raw Data

All raw data including AMT evaluations are publicly available at `https://s3.amazonaws.com/chatbot-eval-data/index.html`. For ease of analysis, the data is also available in a MySQL database hosted on Google Cloud Engine as well as in JSON file format. A template analysis script Python Notebook is available in our repository and also on Google Colab. The ChatEval dataset is potentially useful for the creation and evaluation of automatic metrics.

Selection of Baselines

We seek to establish reasonable public baselines for Seq2Seq-based chatbots. All models trained by us use the OpenNMT-py (Klein et al., 2017) Seq2Seq implementation with its default parameters: two layers of LSTMs with 512 hidden neurons for the bidirectional encoder and the unidirectional decoder. We trained models on three standard datasets: OpenSubtitles, SubTle, and Twitter, and plan to introduce baselines trained on other datasets.

The number of baseline methods will continue to grow. We plan to add an information retrieval baseline, the hierarchical encoder-decoder model (Serban et al., 2017), and several other baselines from ParlAI.

Conclusion and Future Work

ChatEval is a framework for systematic evaluation of chatbots. The ChatEval website provides a repository of model code and parameters, evaluation sets, model comparisons, and a standard

[12] `https://github.com/chateval/chateval`

human evaluation setup. ChatEval seamlessly allows researchers to make systematic and consistent comparisons of conversational agents. We hope that future researchers–and the entire field–will benefit from ChatEval.

Future work includes optional larger evaluation sets for automatic descriptive metrics, such as lexical diversity (*distinct-n*), as these methods are often better suited for larger datasets.

We also plan to extend the ChatEval framework to further tasks by creating multiple new websites (IREval for information retreval, TaskEval for task-based chatbot evaluation, and NLGEval for natural language generation). Each of these will be specialized with small changes to the common framework for the different tasks.

Acknowledgements

We would like to thank Marianna Apidianaki for her helpful feedback on ChatEval and Claire Daniele for proofreading. We thank Sander Wubben and Flow.ai for the helpful API and hosting our interactive baseline session. Finally, we thank the anonymous reviewers for their feedback.

This work was partially supported by João Sedoc's Microsoft Research Dissertation Grant. Thank you to all of the workers on Amazon Mechanical Turk who contribute to our system.

References

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3970–3980. Association for Computational Linguistics.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213–220.

Susan E. Embretson and Steven P. Reise. 2013. *Item response theory*. Psychology Press.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Antske Fokkens, Marieke Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1691–1701.

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.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *ACL, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Kyusong Lee, Tiancheng Zhao, Alan W Black, and Maxine Eskenazi. 2018. Dialcrowd: A toolkit for easy dialog system assessment. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 245–248. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2019. Dialogue generation: From imitation learning to inverse reinforcement learning. In *Thirty-Third AAAI Conference on Artificial Intelligence*.

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 *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. 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 *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Naoki Otani, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2016. Irt-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 511–520. Association for Computational Linguistics.

Iulian V. Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 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. In *NIPS 2017 Conversational AI workshop*.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

Sander Wubben. 2018. Flowai-nlg-api. https://github.com/flow-ai/flowai-nlg-api.

Zhou Yu, Alexandros Papangelis, and Alexander Rudnicky. 2015. Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness. In *2015 AAAI Spring symposium series*.

Steven Kester Yuwono, Wu Biao, and Luis Fernando DHaro. 2018. Automated scoring of chatbot responses in conversational dialogue. In *Third Workshop on Chatbots and Conversational Agent Technologies*.

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.