ChatEval: A Tool for Chatbot Evaluation

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

Abstract

Open domain dialog systems are difficult to evaluate. The current best practice for analyzing and comparing open-domain dialog systems is the use of human judgments. However, the lack of standardization in training and evaluation procedures, and the fact that model parameters and code are rarely published, hinder systematic human evaluation experiments. We propose a unified framework for human evaluation of chatbots that addresses these concerns. ChatEval consists of a web interface for sharing and evaluating chatbots. Researchers can submit their trained models to the ChatEval web interface to effortlessly receive back comparisons with baselines and prior work. Since all evaluation code is open-source, we ensure evaluation is performed in a standardized and transparent way. In addition, we introduce open-source baseline models and several public evaluation sets. ChatEval can be found at http://chateval.org. There is a video demo available at http://chateval.org/video.html.

1 Introduction

Reproducibility and model assessment for dialog systems is challenging, as many small variations in the training setup or evaluation technique can result in large differences in perceived model performance. There are three essential aspects to developing a novel conversational agent: the dataset, the method, and the evaluation scheme. Most papers focus on new methods; however, insufficient attention is paid to ensuring that datasets and evaluation remain consistent and reproducible. For example, we lack standardized dataset splits into train, development, test, and evaluation sets for rigorous model evaluation. Furthermore, most existing work provides insufficient detail to reproduce the method or even compare results. 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. This is especially problematic since deep learning models can take roughly a month to suitably train (Li et al., 2016). However, *ChatEval is not exclusively for neural chatbot*, information retrieval and rule based chatbots can also be evaluated.

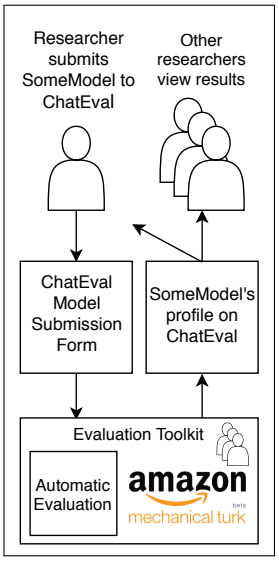

Figure 1: Flow of information in ChatEval. A researcher submits information about her model, including its responses to prompts in a standard evaluation set. Automatic evaluation as well as human evaluation are conducted, then the results are posted publicly on the ChatEval website.

As the field of neural chatbots has grown, it has become increasingly fragmented. 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). Because the NCM is 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. Two such examples can be seen in Table 3.

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 using the same evaluation dataset or experiment procedure (see Table 1).

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 grow with participation

In the ChatEval framework, researchers submit available code, model parameters, and their model's responses to a standardized test set. The ChatEval framework automatically launches the automatic evaluation metrics; however, the ChatEval team manually checks and then launches a human evaluation experiment, comparing the new method against previously submitted approaches and selected baselines. When evaluation is complete, a public profile page for the model is added to the ChatEval website, linking to all relevant information about the model.

Our major contributions are as follows:

- We provide a framework, ChatEval, for researchers to transparently publish their code and effortlessly receive feedback with which to validate methodological improvements.
- We develop credible public baselines using OpenNMT, a well-established open-source framework for machine translation.
- We propose standardized datasets for human evaluation that mimic the actual conversations humans might have with a chatbot.

2 Related Work

Since the original NCM paper was published, many have proposed algorithmic improvements to the basic Seq2Seq approach. Nearly all report some sort of human evaluation results. A sampling of these are shown in Table 1. It is clear that there is little consensus on how to conduct human evaluation.

There has been some work to diminish reliance on human evaluation by coming up with automatic metrics that correlate well with human ratings (Tao et al., 2017; Lowe et al., 2017). More work needs to be done in this area before it is an adequate replacement for human evaluation.

Competitions such as the Alexa Prize[1], ConvAI[2] and WOCHAT,[3] rank submitted chatbots by having humans converse with them then rate the quality of the conversation. However, asking for absolute assessments of quality yields less discriminative results. In the dataset introduced for the ConvAI2 competition, nearly all the proposed methods 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 our tool is entirely focuses on the evaluation and warehousing of models. We do not require any integration of a code base with our infrastructure. This avoids friction in terms of model analysis. ChatEval is a complementary system to ParlAI.

RankMe[5] (Novikova et al., 2018) is an evaluation system for natural language generation. However, although RankMe could be adapted for dialog, this would require substantial work. Furthermore, RankMe does not warehouse evaluation sets and model parameters.

3 The ChatEval Web Interface

The ChatEval web interface consists of four primary pages. They are discussed in the following sections.

[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

Paper	Eval Dataset	# Prompts	# Annotators
(Vinyals and Le, 2015)	Handpicked	200	4
(Sordoni et al., 2015)	Twitter, entire test set	2114	5
(Li et al., 2017)	OpenSubtitles test set, randomly sampled	200	5
(Zhou et al., 2017)	STC test set, randomly sampled	200	no info
(Ghazvininejad et al., 2018)	Not specified	500	7

Table 1: Two-choice human evaluation schemes used by a representative selection of work in the field. Variations in task description are not included here. For example, Ghazvininejad et al. (2018) had evaluators rate separately on "appropriateness" and "informativeness," while others simply asked "Which is better?"

3.1 Home Page

The home page describes the available evaluation sets and how baseline models were chosen. It also links to this paper.

3.2 Model Submission

When researchers submit their model for evaluation, they are also asked to submit the following:

- Description of model which could include link to paper or project page.
- Model's responses on at least one of our evaluation datasets.
- Optionally a URL to a public code repository.
- Optionally a URL to download trained model parameters.

After the code and model parameters are manually checked, we use the ChatEval evaluation toolkit (Section 4) to launch evaluation on the submitted responses. On multiple submissions of a model we ensure that the new submission is adequately different using Jensen-Shannon divergence (Dagan et al., 1997) before launching manual evaluation. 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.

3.3 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.

3.4 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.

4 Evaluation Toolkit

As shown in Figure 1, the ChatEval evaluation toolkit is used to evaluate submitted models. It consists of an automatic evaluation and a human evaluation component.

4.1 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).
- Hits@1 which is the ratio of times that the model chooses the correct response relative to random distractor responses (Zhang et al., 2018).

Our system is easily extensible to support other evaluation metrics, and we plan to expand this list.

4.2 Human Evaluation

Two-choice tests consist of showing the evaluator a prompt and two possible responses. 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

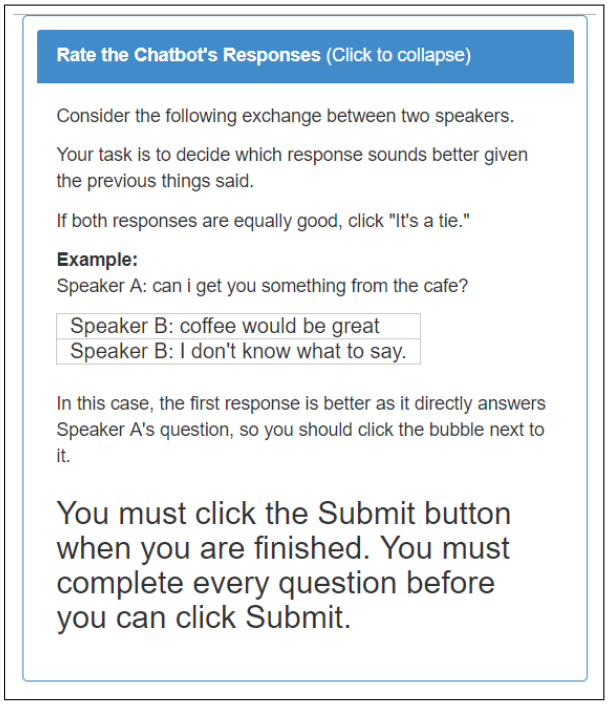

Figure 2: The instructions seen by AMT workers.

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 prompt, and on average it takes 1 minute to evaluate 10 choices with a maximum allowed time of 2 minutes. We used three evaluators per prompt, so, if there are 200 prompts, 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. Unfortunately, there are occasionally bad workers, which we remove from our results. In order to identify such workers, we examine the worker against the other two annotators. The overall IAA varies between .2 to .54 if we include tie choices. However, we only care about win or loss choices with this correction the IAA as measured by Cohen's weighted kappa (Cohen, 1968) is .55. This is likely sufficient for assessing and training automatic metrics.

4.3 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.[6] We provide clear instructions for researchers to perform human and automatic evaluation on their own with the toolkit as an alternative to using our web interface. Furthermore, we plan to integrate the AMT interface with ParlAI, to extend the existing evaluation in ParlAI and allow users to evaluate easily.

4.4 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 [6] and also on Google Colab. The ChatEval dataset is potentially useful for the creation and evaluation of automatic metrics.

5 Evaluation Datasets

When choosing evaluation datasets to include in ChatEval, we considered the fact that even non-task driven chatbots have a goal. Perhaps one's purpose is text messaging or another's aim is to engage in voice conversations. Many existing works fail to take into account their chatbot's goal when picking prompts for human evaluation. Randomly sampling from a test set with the same or similar distribution to the training set can create a mismatch between the distribution of the train/test set and the actual types of conversation we would expect a user to have when chatting. This is very much true for models trained on OpenSubtitles (Asghar et al., 2017), where the end goal is likely not to generate movie scripts. For models trained on Twitter, it is reasonable to form a test set out of Tweets only if the aim is to have a chatbot that responds well to Tweets.

In ChatEval, we prioritize the goal of engaging in text-based interactions with users who know they are speaking with a chatbot. 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,

[6] https://github.com/chateval/chateval

Train dataset	Eval Ppl	Distinct-1	Distinct-2	Avg len	Human 1 Sim	Human 2 Sim
OpenNMT	**3.89**	.12	.16	3.43	.401	.386
CakeChat	-	**.18**	**.41**	**8.48**	**.446**	**.426**
Twitter 2016 (ours)	7.55	.12	.24	5.93	.403	.377
OS (ours)	4.19	**.18**	.19	2.12	.330	.348
OS Q (ours)	4.57	.14	.27	3.63	.422	.435

Table 2: Automatic evaluation metrics computed in the ChatEval system on the NCM evaluation dataset. All models are Seq2Seq models with attention. OS Q is a filtered version of OpenSubtitles containing only questions. OpenNMT is a pre-trained benchmark model released by OpenNMT and trained on OpenSubtitles. Sim is the cosine-similarity between mean word embeddings of predicted and ground-truth responses.

Iris, and CIC. Participants knew that they were speaking with a chatbot, and the conversations reflect this. 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.

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. This was necessary in order to compare our models' performance with the NCM, since only the NCM's responses, and not its source code, have been released publicly.

ChatEval can be trivially extended to support other evaluation datasets, and the number of supported datasets to grow. While our first batch of released datasets all have only a single conversation turn as a prompt, our framework can also support datasets with multi-turn prompts, and we plan to add these in the future.

6 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 and also incorporate ParlAI baseline models.

The baselines currently included in ChatEval are shown in Table 2. In addition to our own trained models, we also include the OpenNMT benchmark for dialog systems[7] and a model released by Cakechat[8], an open-source system from the chatbot start-up Luka Labs. A comparison of these baselines is in Table 3.

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., 2016), and several other baselines from ParlAI.

Resp 1	Resp 2	Gain
NCM	OpenNMT	.470 +/- .060
NCM	Cakechat	.285 +/- .013
OpenNMT	Cakechat	.039 +/- .076

Table 3: Comparison of NCM vs OpenNMT Seq2Seq model, NCM vs Cakechat and OpenNMT vs Cakechat.

6.1 Case Study: Effect of Training Dataset Filtering

We used ChatEval to understand the effect of data filtering on model preference. In particular, we tested the effect of this on model performance of training the Seq2Seq model on (1) filtering on the Twitter dataset and (2) using only a subset from the OpenSubtitles dataset. For both Twitter and OpenSubtitles, we used the ParlAI framework (Miller et al., 2017) to generate the train, validation, and test sets.

Since Twitter is an extremely noisy dataset, most papers perform some sort of cleanup following tokenization. We applied "garbage removal", similar to the method from Ritter et al. (2010). We removed company-based conversations and Tweets with profanity. In total roughly 35,000 out of 2.5 million tweets were removed. Using ChatEval we found that AMT raters had no statistically significant preference.

[7] http://opennmt.net/Models-py/
[8] https://github.com/lukalabs/cakechat

For OpenSubtitles, we trained a model only on prompts with a question mark as the last token, which is only 15% of the original dataset. A model (OSQ) trained on only the subset performed better on the NCM evaluation set than did the model trained on the entire dataset. In part this is due to the fact that the evaluation set is heavily weighted towards interrogative sentences. However, the 13.5% gain is quite striking and equivalent to some reported model architecture changes (Asghar et al., 2017). We also found similar gains using the DBDC evaluation set.

Our case study is illustrative as it shows how ChatEval can help researchers to gain insights into what kind of standardization might be possible and important.

7 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

Nabiha Asghar, Pascal Poupart, Xin Jiang, and Hang Li. 2017. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 78–83. Association for Computational Linguistics.

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

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *EACL*, pages 56–63. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2017. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Dialog System Technology Challenge 6 (DSTC6) Workshop*.

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.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *ACL*, pages 994–1003. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Data Distillation for Controlling Specificity in Dialogue Generation.

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.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *ACL*, pages 1116–1126. 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.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *NAACL*, New Orleans, Louisiana.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *ACL*, pages 172–180.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses.

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

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems.

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.

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.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory.