ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons

Abstract

While dialogue remains an important end-goal of natural language research, the difficulty of evaluation is an oft-quoted reason why it remains troublesome to make real progress towards its solution. Evaluation difficulties are actually two-fold: not only do automatic metrics not correlate well with human judgments, but also human judgments themselves are in fact difficult to measure. The two most used human judgment tests, single-turn pairwise evaluation and multi-turn Likert scores, both have serious flaws as we discuss in this work. ::: We instead provide a novel procedure involving comparing two full dialogues, where a human judge is asked to pay attention to only one speaker within each, and make a pairwise judgment. The questions themselves are optimized to maximize the robustness of judgments across different annotators, resulting in better tests. We also show how these tests work in self-play model chat setups, resulting in faster, cheaper tests. We hope these tests become the de facto standard, and will release open-source code to that end.

Introduction

Dialogue between human and machine is an important end-goal of natural language research. The open-ended nature of generating sequences in a multi-turn setup naturally makes the task difficult to evaluate – with full evaluation possessing many of the difficulties of the task itself as it requires deep understanding of the content of the conversation. As in many other natural language generation (NLG) tasks, automatic metrics have not been shown to have a clear correlation with human evaluations BIBREF0, BIBREF1. This means the current standard for all dialogue research involves human trials, which slows down research and greatly increases the cost of model development.

Unfortunately, human judgments are themselves difficult to measure. The two most used approaches, single-turn pairwise evaluation BIBREF2, BIBREF3, and multi-turn Likert scores BIBREF4, BIBREF5, BIBREF6, BIBREF7, BIBREF8 have serious limitations. Single-turn pairwise evaluation provides the benefits and simplicity of an A/B test, allowing for cheap and fast annotations, with comparisons that are robust to annotator score bias, but fail to take into account the multi-turn aspect of conversations. To give a trivial example, such comparisons fail to capture whether the model would repeat itself in a multi-turn conversation because they only look at one turn; repetition is a known issue that humans dislike BIBREF6.

Multi-turn Likert scores require the annotator to have a multi-turn conversation and then provide an integer score, which is more costly and time-consuming to run but evaluates full conversations more accurately. The integer scores however suffer from differing bias and variance per annotator, which researchers have tried to mitigate BIBREF9, but nevertheless due to its lack of sensitivity often yields comparisons that are not statistically significant. Furthermore, due to strong anchoring effects during model evaluation, i.e. that annotators are affected by the first systems they evaluate, Likert comparisons are generally not comparable across multiple papers. This mandates that evaluations of new models be simultaneously collected with baselines, further increasing the cost of developing additional models BIBREF6.

In this work we introduce Acute-eval, a method that combines the benefits, and attempts to mitigate the deficiencies, of the above two approaches by introducing a pairwise relative comparison setup for multi-turn dialogues. In each trial, we show the annotator two whole conversations, with the second speaker in each conversation highlighted, as the judgment should be independent of the quality of the first speaker, see Figure FIGREF1. We then show a carefully worded question with two choices: speaker A or B, where the question measures a desired quality such as which speaker is more engaging, interesting or knowledgeable. Our experiments show that annotators perform well in this setup, and that our method can reveal subtle but significant differences between conversational models that other approaches, such as multi-turn Likert, cannot.

Overall, our work provides the following contributions:

A new evaluation method with a clear mechanism that provides fast, cheap iteration. This evaluation method allows efficient reuse of data from prior papers, allowing new models to be evaluated independently of baselines, and dramatically lowers the cost of annotation.

We optimize question choices to find those with the highest agreement, increasing confidence in the desired test. We provide the wording of the questions that we found to work best for several questions of interest (most engaging, human, interesting or knowledgeable conversationalist) for further research use.

We provide an explicit benchmark comparison between current best performing retrieval and generative models on two recent tasks, PersonaChat BIBREF5 and Wizard of Wikipedia BIBREF7 for several question choices, revealing the current state-of-the-art, and to be used for benchmarking on these tasks in the future.

We show that our test can be applied to self-chats rather than human-model conversation logs, which can reveal problems with existing models at a cheaper price, and provides high agreement with the human-model evaluations.

We will release the code for running these tests.

Related Work

Dialogue tasks have traditionally been separated into two areas: goal-oriented and chitchat. Goal-oriented tasks typically have a clearer evaluation, e.g. task completion can be measured if the correct actions are taken BIBREF10, BIBREF11, BIBREF12, BIBREF13, BIBREF14. Chitchat tasks are more open ended, and instead feature conversations without a precise goal that can be automatically evaluated. For example, conversations where two speaking partners are discussing interests BIBREF5 or topics BIBREF7. We study the latter in this work.

Evaluation of chitchat tasks with automatic metrics is difficult precisely because of their open-ended nature. For example, the answer to the question “What are you doing tonight?” has many possible answers, each with little word overlap. This means standard metrics for tasks like question-answering or machine translation do not work well, and have poor correlation with human judgments BIBREF0, BIBREF15. Nevertheless, a number of studies do report automatic metrics, without human studies BIBREF16, BIBREF17. Researchers have made attempts to improve automatic evaluation, trying methods such as adversarial evaluation BIBREF18, learning a scoring model BIBREF1, or a learnt ensemble of automatic metrics BIBREF19, but their value is as yet not fully understood.

Currently the standard approach in chitchat dialogue is to perform human evaluations BIBREF2, BIBREF20, BIBREF21, BIBREF4, BIBREF5, BIBREF7, typically reporting a judgment such as conversation quality or appropriateness via a Likert scale or pairwise comparison. While conversations are naturally multi-turn, pairwise setups typically consider single turn evaluations, taking the “gold” dialogue history from human-human logs, and only consider altering a single utterance. A more complete multi-turn evaluation is typically measured with a Likert scale (usually 1-4 or 1-5) after the conversation takes place. Some works such as BIBREF6 ask a series of questions relating to different aspects of conversational ability. There are some notable variants from these standard setups. BIBREF22 provide a method that combines continuous scales and relative assessments, but in single-turn, rather than multi-turn evaluation. BIBREF19 compare human evaluations to automatic metrics computed on self-chats. Note that we also use self-chats in this work, but we evaluate these with humans, rather than automatic metrics.

Finally, this work expands upon some of the ideas present in BIBREF6. In that work, a test for interestingness of a specificity-controlled model conducted with pairwise chat logs was mentioned, similar to the ones used here, but was not the focus of their work. In our work, we conduct a full study of novel variants of this approach, consider optimizing the questions for robust measurements over four types of questions, utilize self-chat logs in addition to human-bot logs, and benchmark state-of-the-art models across two recent tasks.

Method: Acute-eval

To compare two dialogue models, model A and model B, our evaluation asks humans to directly compare side-by-side multi-turn dialogues conducted by these models. See Figure FIGREF1 for an example.

Our method is thus the following: (1) collect conversation logs for model A; similarly for model B. (2) In a number of trials, ask annotators to make binary judgments between sampled pairs from the logs, and collate the results to determine the winner, either A or B, and the statistical significance.

We consider different approaches to step (1) and (2) below.

Method: Acute-eval ::: Human-Model chats

Our standard setup is to compare conversation logs between models and humans. In each evaluation trial we then show a human annotator two of the previously obtained conversations, one of model A conversing with a human, and one of model B conversing with a (possibly different) human. The annotator sees the conversations side by side on the same screen, with the two models' utterances highlighted in different colors, and the human utterances in gray to minimally distract from the models.

The annotator is posed a question phrasing (e.g. “which speaker is more knowledgeable” or “which speaker sounds more human?”), and asked to make a binary choice between model A and model B. They are strongly encouraged to provide a short text justification for their choice. We collect N trials of such pairwise judgments, and use them to decide which model wins. Statistical significance can be computed using a binomial test.

Method: Acute-eval ::: Self-Chats

Human-model conversation logs are themselves time-consuming and expensive to collect, which limits rapid iterative model development. We investigate if it is possible to remove the human from the conversation, and only use human annotators in the final pairwise conversation evaluation step. The concept of self-chats BIBREF21, BIBREF19, whereby a model talks to itself, playing the roles of both speaking partners, has been previously explored in other contexts. Such logs are easy to collect for models A and B, involving simply running inference for both speaker roles. We then use these logs in the Acute-eval pairwise comparison setup as described above.

Method: Acute-eval ::: Question Optimization

So far, we have not detailed the actual question(s) asked of the annotators. The framing and phrasing of questions in surveys is known to greatly affect the direction of responses, and therefore, in the case of evaluation, inter-annotator agreement. Though this has been noted in prior work BIBREF1, we have found no systematic experimentation on question formulation or task presentation. We therefore aim to propose and evaluate multiple potential question wordings to achieve higher agreement.

To do this, we build an initial test that compares human-human logs with human-model logs where the model is a relatively low quality baseline model. The aim is that there should be a clear and agreeable difference between human and model which is visible to human annotators. We ask annotators to make judgments between these two, where we choose pairs where the human should be judged as superior.

We then run independent trials with different question phrasing, and find the questions with highest inter-annotator agreement. The winning questions can then be used in future experiments by ourselves, and other researchers. Although having high inter-annotator agreement does not guarantee that crowdworkers interpret the question as intended, it increases the chance the question is understood uniformly. That is, the researcher still has to exercise care in the formulation of the question so that they believe it measures the quantity they are interested in. In our experiments we find questions with high-agreement rate over four axes: engagingness, interestingness, knowledge and humanness.

Method: Acute-eval ::: Annotation Quality

We use crowdworkers for our annotations. We recommend limiting the number of annotations a single worker may complete to be only a few pairs (in our experiments, if we are making N model comparisons then we allow N annotations). In preliminary trials, we found that limiting the influence of any one worker was important for replicability, but that results were highly consistent across multiple runs with this limitation.

Additionally, the first comparison any worker is asked to annotate consists of a conversation between a weak baseline model and human, and a human-human conversation. If a worker fails to rate the human-human conversation as better, we remove their annotations from the results, in order to remove poor quality annotators. We additionally remove workers who never give a reason for their choice. Note that adding such worker quality tests to pairwise annotation tasks is straightforward where the gold annotation is known, while it is harder for Likert tests which have integer scores. One may also increase the number of quality-control annotations to decrease the likelihood of fraudulent workers, but we found using a single control question had a reasonable cost-noise ratio.

Each specific pair of conversations is shown at most once, given that there are at least as many possible pairs of conversations as desired annotations. If there are more conversations available for each model than desired annotations, each conversation is shown at most once - that is, in only one annotation. We found that maximizing the diversity of pairs improved robustness of our evaluation across multiple replication experiments.

Experiments

We perform experiments on two tasks, PersonaChat and Wizard of Wikipedia, which evaluate different aspects of conversational ability. We first optimize the questions to maximize worker agreement, and then benchmark existing state-of-the-art models on each task.

Experiments ::: PersonaChat task

PersonaChat BIBREF5 is a chitchat dialogue task involving two participants (two humans or a human and a bot). Each participant is given a persona – a short collection of personal traits such as I'm left handed or My favorite season is spring – and are instructed to get to know each other by chatting naturally using their designated personas, for 6–8 turns. The original dataset contains nearly 9000 human-human training conversations; most models are pretrained with a larger corpus, and then fine-tuned on this set.

PersonaChat was the subject of the NeurIPS 2018 ConvAI2 Challenge BIBREF8, in which competitor's models were first evaluated with respect to automatic metrics, and then with respect to human judgment via human-bot chats followed by the question “How much did you enjoy talking to this user?" on a scale of 1–4. A total of 9 systems were evaluated using human annotators, 100 conversations for each. In this work, we leverage the human-model chat logs from the ConvAI2 competition for three models: Lost in Conversation (LIC), which won the competition, and Hugging Face (HF; BIBREF23, BIBREF23) which won the automatic evaluation track, and the KVMemNN BIBREF24 baseline released by the competition organizers (KV; BIBREF8, BIBREF8). LIC and HF are large pretrained and fine-tuned generative Transformer models, while KV is a retrieval model with no pretraining.

Secondly, we also compare to recently published models from BIBREF6. The authors studied the effects of controllable generation. and showed that Repetition-controlled (RC), Inquisitive (INQ), and Interesting (INT) models obtained the highest human Likert scores in their study, however their comparison to models from other studies is not direct. We thus compare to these models as well; we use the human-model conversation logs from their work, 100 for each model.

Finally, we also compare to the Polyencoder model (PE, BIBREF25, BIBREF25), a recent state-of-the-art retrieval model. It is a type of large Transformer architecture pretrained on Reddit, which learns a small number of global features to represent the input so that retrieval can be computed efficiently. As no conversation logs were provided in that work, we additionally collect human-model conversations for that model.

Overall, we benchmark 7 models, and compare them to human (H) performance in a number of different settings: with human-model and self-chat over three questions: engagingness, humamnness and interestingness.

Experiments ::: Wizard of Wikipedia task

Wizard of Wikipedia BIBREF7 is a chitchat dialogue task where two speakers discuss a topic in depth, chosen from 1247 topics. One speaker (termed the Wizard) is meant to be both engaging and knowledgeable on the topics, and has access to an information retrieval system over Wikipedia to supplement their own knowledge. The other speaker (the Apprentice) is meant to be curious and eager to learn about the topic. The original dataset contains over 18,000 human-human dialogues, and has been used to train various kinds of models to imitate the human wizards. These include the Memory Network Transformer, in both generative and retrieval versions that employs the retrieved knowledge by attending over it before producing an utterance (GK and RK respectively), and baselines that do not have access to the knowledge (GU and RU). See Figure FIGREF25 for an example chat. We use the human-model logs from that paper (100 conversations for each model) on unseen test topics and evaluate them against humans (H), using both engagingness and knowledgeability questions. We note the original paper tested engagingness only.

Experiments ::: Question Optimization

We are interested in evaluating models in terms of four axes: engagingness, interestingness, knowledge and humanness. In order to find the questions with highest inter-annotator agreement, we run multiple trials of experiments according to the setup described below. Each trial tests the effectiveness of a single question and consists of the same set of multi-turn conversation logs, presented to the human annotators. We test 13 questions: three regarding engagingness, four regarding interestingness, three regarding humanness, and three regarding knowledgeability (see Table TABREF11).

We compare human-human logs with human-model logs where the model is a relatively low quality baseline model, with the aim that there should be a clear and agreeable difference between human and model which is visible to human annotators. For PersonaChat we use a greedy generative baseline, and for Wizard we use the GU (generative unknowledgeable) model. Both of these baselines exhibit strong repetitive behavior which is known to be highly disfavored by crowdworkers BIBREF6. We select a single handpicked conversation pair for each of the tasks, and collect $\sim $20 annotations per question.

We calculate the inter-annotator agreement for each question. The question achieving the highest inter-annotator agreement is selected for use in the rest of our experiments. The specific question phrasing and the texts accompanying the option for Speaker 1 (i.e. the left-hand conversation) are listed in Table TABREF11 along with inter-annotator agreements. As can be seen, the phrasing of the question is important, with poor phrasing choices leading to much lower agreement levels, e.g. 86.7% agreement in the best case for interestingness, and 69.6% in the worst case.

As a preliminary sanity check, we ran A/A tests over each of the engagingness, interestingness, and humanness best questions, with the same model appearing as both Speaker 1 and 2. All three tests came back close to 50-50.

Overall, we see this question optimization step as an important pre-requisite for our main experiments, and use the best discovered phrasing in each case. We encourage further research to use them as well.

Experiments ::: Benchmarking: Evaluation of State-of-the-art ::: PersonaChat

We first compare all 7 models and humans on the PersonaChat task using Acute-eval over the human-model chats using the optimized engagingness question. In total, we evaluate 28 paired comparisons. Results are given in Table TABREF18. Bold win percentages indicate significance.

We first observe that the models form a clean well-ordered set, and there are no rock-paper-scissors effects, giving an order Human $>$ PE $>$ LIC $>$ INT $>$ HF $>$ INQ $>$ KV $>$ RC. In general, these results agree closely with the known Likert comparisons made in prior papers, shown in Table

TABREF19. Similar conclusions are derived for the interestingness and humanness questions as well, see Tables TABREF26 and TABREF24, note the model ordering is slightly different for those questions. BIBREF6 previously showed that different models often exhibit different rankings for different metrics, and Acute-eval results remain largely consistent with Likert.

A surprising result for the community is that the retrieval model PE outperforms all generative models, as the community has focused heavily on building generative models, e.g. almost all 23 entrants to the ConvAI2 competition BIBREF8. Now that the current best performing models have been benchmarked against each other we hope future research will use the same approach so the state-of-the-art can be clearly tracked.

Experiments ::: Benchmarking: Evaluation of State-of-the-art ::: Self-Chat

We perform Acute-eval over self-chats instead of human-model chats. We compare all models and humans (via human-human chats) in an otherwise identical setup to the human-bot evaluation for PersonaChat. Results are given in Table TABREF20.

We observe very similar conclusions to human-model chats in terms of winning models, making this a viable cheaper alternative to collecting human-model conversations, thus being considerably cheaper to collect. This approach also appears to require relatively fewer annotations/person-hours in this case to achieve statistical significance. One important caveat is the performance of the HF model. HF self-chats surface degeneracies in the model itself, and do not look natural (see Figure FIGREF22 for examples), explaining its poor performance compared to all other models. All other models do not exhibit this behavior and apart from HF, are ordered by humans exactly the same as for human-bot chats. For example, see Figure FIGREF23 for PE engaging in self-chat more successfully. However, due to the inadequacies of a specific model, in this case HF, conclusions from self-chat performance results must therefore be handled with care, but we believe are a reasonable choice for early experiments in the model development cycle, enabling faster research iteration.

One concern with self-chat is that powerful models could easily cheat, and simply recall training examples with perfect accuracy. In practice, we found that none of the models exhibit this behavior: $<$1% of the Polyencoder's call-response utterance pairs produced during self-chats come directly from the training set. The worst offender, INQ, has roughly 10% of pairs coming from training, but this stems from it using the same generic greeting and response in nearly all conversations (“Hello, how are you doing today?”, “I am doing well, how about yourself?”).

Experiments ::: Benchmarking: Evaluation of State-of-the-art ::: Wizard of Wikipedia

We similarly compare all 4 models and humans on the optimized engaging and knowledge questions. The results are given in Tables TABREF27 and TABREF28. We again find retrieval models outperform generative models, with knowledge attention (GK) clearly helping the generative models, but with RU and RK very close.

Results largely agree between the two questions, except retrieval with knowledge (RK) more clearly beats the generative version (GK) than retrieval without (RU) when the question is about knowledge. For the engagingness question, where it makes sense that this is less important, there is little difference between knowledge or not.

Experiments ::: Benchmarking: Evaluation of State-of-the-art ::: Comparison to Likert

We compare Acute-eval to multi-turn Likert for both tasks by computing pairwise Likert differences, where known, from the original papers. We do not compare across papers as evaluation setups differ. Values are provided in Tables TABREF19, TABREF26, TABREF24 and TABREF27. While the tests generally agree, Acute-eval can be a more sensitive test, which more often yields significance. On Wizard of Wikipedia where all Likert matchups are known, 8 of the pairwise matchups are significant for our test with human-model chats, while 6 are significant for Likert. On PersonaChat for the interestingness question, 6 of 10 matchups are significant for Acute-eval, including all known Likert matchups, which only has 2 of 3 that are significant. For the humanness question, 5 of 10 matchups are significant for Acute-eval, including all known Likert matchups, which only has 2 of 3 that are significant. For the engagingness question, 5 of the 9 Likert matchups are significant. All 9 are significant for Acute-eval when using self-chats; 3 are significant for human-model chats.

We compare the cost effectiveness of Likert to Acute-eval human-model and self-chat comparisons in Figure FIGREF30. Shown is the PersonaChat Engagingness question comparing RC and INT models, a fairly tight matchup. We show the % chance of achieving significance when drawing pairs of dialogues at random, plotting with respect to person-hours spent annotating. In this case Likert fails to achieve significance, likely due to bias and variance issues with integer scores. Acute-eval human-model and self-chat pairwise tests perform well, achieving significance; self-chat requires fewer person-hours.

Conclusion

Studying the ability of machines to communicate with humans is an important long-term goal of AI research. Unfortunately, measuring progress towards that goal has been hampered by the trustworthiness of evaluation itself. Current human evaluation methods such as multi-turn Likert are expensive to run, have annotator bias and variance problems, and can fail to yield statistical significance.

In this work we have contributed a novel evaluation method that alleviates some of these problems. By optimizing questions and performing comparisons on pairs of human-bot dialogues we arrive at more sensitive statistical tests when benchmarking current state-of-the models. Utilizing self-chat bot evaluations we can often improve sensitivity, while yielding even cheaper evaluations. We will publicly release the code for our tests, and recommend them to be used in future research studies in order to push forward the state of the art.