Ditch the Gold Standard: Re-evaluating Conversational Question Answering

Anonymous ACL submission

Abstract

Conversational question answering (CQA) systems aim to provide natural-language answers to users in information-seeking conversations. Existing benchmarks compare CQA models with pre-collected human-human conversations, using ground-truth answers provided in conversational history. It remains unclear whether we can rely on this static evaluation for model development, and whether current systems can well generalize to real-world human-machine conversations. In this work, we conduct the first large-scale human evaluation of state-of-the-art CQA systems, where human evaluators converse with models and judge the correctness of their answers. We find that the distribution of human-machine conversations drastically differs from that of human-human conversations, and evaluating with gold answers is inconsistent with human evaluation. We further investigate how to improve automatic evaluations and propose a question rewriting mechanism based on predicted history, which better correlates with human judgments. Finally, we analyze the impact of various modeling strategies. We hope our findings can shed light to how to develop better CQA systems in the future.

1 Introduction

Conversational question answering (CQA) aims to build machines to answer questions in conversations, and has the promise to revolutionize the way humans interact with machines for information-seeking. With recent development of large-scale datasets (Choi et al., 2018; Saeidi et al., 2018; Reddy et al., 2019; Campos et al., 2020), rapid progress has been made in better modeling of conversational QA systems.

Current datasets are collected by crowdsourcing human-human conversations, where the questioner asks questions based on an evidence passage and conversational history and the answerer provides corresponding answers. When evaluating CQA systems, a set of held-out conversations are used for asking models questions in turn. Since the evaluation builds on pre-collected conversations, the *gold history* of the conversation is always provided, regardless of models' actual predictions (Figure 1). Despite the extremely competitive performance of current systems on this static evaluation, it is questionable whether this can faithfully reflect models' true performance in real-world applications. To what extent do human-machine conversations deviate from human-human conversations? What will happen if models have no access to ground-truth answers in a conversation?

To answer these questions and better understand the performance of CQA systems, we carry out the first large-scale human evaluation with four state-of-the-art models on the QuAC dataset (Choi et al., 2018), by having human evaluators converse with the models and judge the correctness of their answers. We collected 1,446 human-machine conversations in total, with 15,059 question-answer pairs. Through a careful analysis, we identify a significant distribution shift from human-human conversations and discover a clear inconsistency of model performance between current evaluation protocol and human evaluation.

This finding motivates us to improve the automatic evaluation so that it is better aligned with human evaluation. Mandya et al. (2020); Siblini et al. (2021) identify a similar issue in gold-history evaluation and propose to use models' own predictions for automatic evaluation. However, predicted-history evaluation poses another challenge: since all the questions have been collected beforehand, using predicted history will invalidate some of the questions because of changes in the conversational history (see Figure 1 for an example).

Based on this insight, we propose a *question rewriting* mechanism, which automatically detects and rewrites invalid questions with predicted his-

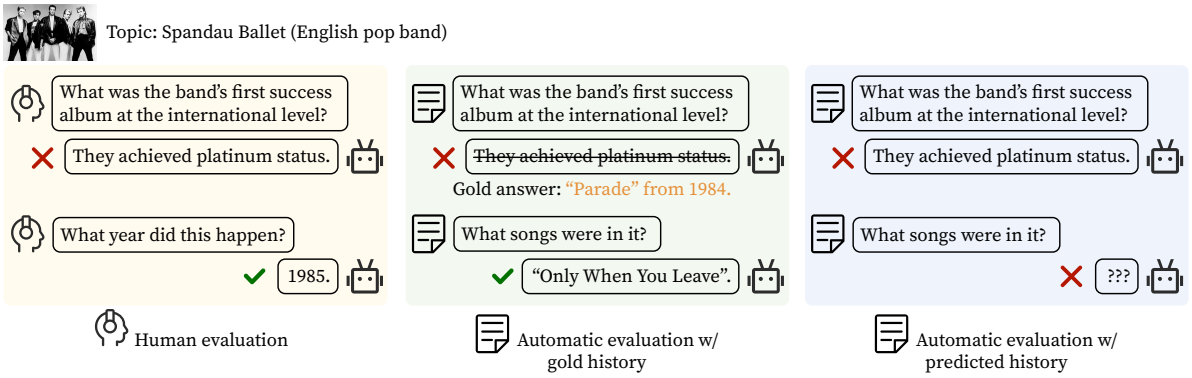


Figure 1: Examples of human evaluation and automatic evaluations with gold history and predicted history. The model answers the first question incorrectly. (a) A human questioner inquires based on the model's prediction. (b) Automatic evaluation with gold history asks pre-written questions and uses gold answers as the conversation history. (c) Using predicted history in automatic evaluation may invalidate the next question.

tory (Figure 4). We use a coreference resolution model (Lee et al., 2018) to detect coreference inconsistency of question text between predicted history and gold history, and then rewrite those questions by substituting with the correct mentions, so that the questions are resolvable in the context. Compared to predicted-history evaluation, we find that incorporating rewriting better aligns with human judgements and reflects models' true performance.

Finally, we also investigate the impact of different modeling strategies on human evaluation. We find that accurately detecting unanswerable questions and explicitly modeling question dependencies on the context are crucial in model performance. Equipped with all the insights, we discuss directions for CQA modeling. We will release our human evaluation dataset and hope that our findings can shed light on future development of better conversational QA systems.

2 Preliminary

2.1 Evaluation of Conversational QA

In conversational question answering, there is an evidence passage P, a (human) questioner $\mathcal{H}$ that has no access to P, and a model $\mathcal{M}$ that has access to P. The questioner asks questions about P and the model answers them based on P and the conversational history so far (see an example in Figure 1). Formally, for the i-th turn, the human asks a question based on the previous conversation,

$$Q_i \sim \mathcal{H}(Q_1, A_1, ..., Q_{i-1}, A_{i-1}),$$

and then the model answers it based on both the history and the passage,

$$A_i \sim \mathcal{M}(P, Q_1, A_1, ..., Q_{i-1}, A_{i-1}, Q_i),$$

where Q_i and A_i represent the question and the answer at the i-th turn. If the question is unanswerable from P, $A_i =$ `CANNOT ANSWER`. The model $\mathcal{M}$ is evaluated by the correctness of answers.

Evaluating CQA systems requires having human in the loop and it is expensive to collect the judgments. Instead, current CQA benchmarks use automatic evaluation with *gold history* (*Auto-Gold*). For example, QuAC (Choi et al., 2018) collects a set of human-human conversations for automatic evaluation. For each passage, one annotator asks questions without seeing the passage, while the other annotator provides the answers. Denote the collected questions and answers as Q_i^* and A_i^* (gold answers). In gold-history evaluation, we inquire the model with pre-defined questions Q_i^*:

$$A_i \sim \mathcal{M}(P, Q_1^*, A_1^*, ..., Q_{i-1}^*, A_{i-1}^*, Q_i^*),$$

and we evaluate the model by comparing A_i to A_i^* (measured by word-level F1). This process does not require human effort, but it can't truly reflect the distribution of real human-machine conversations, since the questioner may ask different questions based on different models' predictions.

In this work, we choose the QuAC dataset for our main evaluation, since it is closer to real-world information-seeking conversations, where the questioner *cannot* see the evidence passage. It prevents the questioner asking questions that simply overlaps with the passage and encourages truly unanswerable questions. QuAC also adopts *extractive* question answering that restricts the answer as a span of text, and more modeling work has been done than in free-form question answering.

2.2 Models

For human evaluation and analysis, we choose the following four representative CQA models:

BERT. A simple BERT baseline, which concatenates the passage, the previous two turns of question-answer pairs, and the question as the input and predicts the answer as in Devlin et al. (2019).[1]

GraphFlow. Chen et al. (2020) propose a recurrent graph neural network on top of BERT embeddings to model the question dependencies on the history and the passage.

HAM. Qu et al. (2019) propose a history attention mechanism (HAM) to softly select the most relevant previous turns.

ExCorD. Kim et al. (2021) train a question rewriting model on CANARD (Elgohary et al., 2019) to generate context-independent questions, and then use both the original and the generated questions to train the QA model. This model achieves the current state-of-the-art on QuAC (67.7 F1).

For all the models except for BERT, we use the original implementations for a direct comparison.

3 Human Evaluation

In this section, we carry out a large-scale human evaluation with the four models discussed above.

3.1 Conversation Collection

We collect human-machine conversations using 100 passages from the QuAC development set on Amazon Mechanical Turk.[2] We also design a set of qualification questions to make sure that the annotators fully understand our annotation guideline. For each model and each passage, we collect three conversations from three different annotators.

We collect each conversation in two steps:

(1) The annotator has no access to the passage and asks questions. The model extracts the answer span from the passage in a human-machine conversation interface.[3] We provide the title, the section title, the background of the passage, and the first question from QuAC as a prompt to annotators. Annotators are required to ask at least 8 and at most 12 questions. We encourage context-dependent questions, but also allow open questions like "What else is interesting" if asking a follow-up question is difficult. (2) After the conversation ends, the annotator is shown the passage and asked to check whether the model's answers are correct or not.

We noticed that the annotators are biased when evaluating the correctness of answers. For questions to which the model answered `CANNOT ANSWER`, annotators tend to mark the answer as incorrect without checking if the question is answerable. Additionally, for answers with the correct types (for example, a date as an answer to "When was it?"), annotators tend to mark it as "correct" without verifying from the passage. Therefore, we asked another group of annotators to verify question answerability and answer correctness.

3.2 Answer Validation

For each collected conversation, we ask two additional annotators to validate annotated answers. First, each annotator reads the passage before seeing the conversation. Then, the annotator sees the question (and question only) and selects whether the question is (a) ungrammatical, (b) unanswerable, or (c) answerable. If the annotator chooses "answerable", the interface then reveals the answer and asks about its correctness. If the answer is "incorrect", the annotator selects the answer span from the passage. We discard all questions that both annotators find "ungrammatical" and the correctness is taken as the majority of the 3 annotations.

In total, we collected 1,446 human-machine conversations and 15,059 question-answer pairs. The data distribution is very different from the human-human conversations (the QuAC dataset): we see more open questions and unanswerable questions, due to less fluent conversation flow caused by model mistakes, and that models cannot provide feedback to questioners like human annotators do. (see more detailed analysis in §6.2). This collection not only supports our comparison and analysis, but also complements existing datasets.

3.3 Annotator Agreement

Deciding the correctness of answers is challenging for humans in some cases, especially when questions are relatively short and ambiguous. We measure annotators' agreement and calculate the Fleiss' Kappa (Fleiss, 1971) on the agreement between annotators in the validation phase. For deciding one turn is unanswerable, correct, or incorrect, we achieve $\kappa = 0.598$ (moderate agreement). For deciding whether one turn is unanswerable, we have

[1] We use `bert-base-uncased` as the encoder.

[2] We restrict the annotators from English-speaking countries, having finished at least 1,000 HITS with an acceptance rate at least 95% for high quality.

[3] We used ParlAI (Miller et al., 2017) to build the interface.

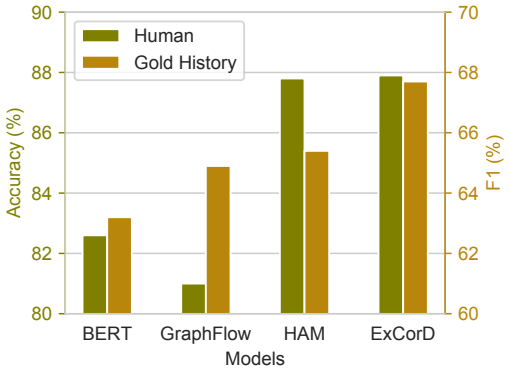


Figure 2: Model performance of human evaluation (accuracy) and automatic evaluation with gold history (F1). Human evaluation and Auto-Gold rank BERT and GraphFlow differently.

Unresolved coreference (44.0%)	
Q_1^*:	What was Frenzal Rhomb's first song?
A_1^*:	Punch in the Face.
A_1:	CANNOT ANSWER.
Q_2^*:	How did it fare?
Incoherence (39.1%)	
Q_1^*:	Did Billy Graham succeed in becoming a chaplain?
A_1^*:	He contracted mumps shortly after...
A_1:	After a period of recuperation in Florida, he ...
Q_2^*:	Did he retire after his mumps diagnosis?
Correct answer changed (16.9%)	
Q_1^*:	Are there any other interesting aspects?
A_1^*:	... Steve Di Giorgio returned to the band...
A_1:	... bassist Greg Christian had left Testament again...
Q_2^*:	What happened following this change in crew?

Figure 3: Examples of invalid questions with predicted history. Some are shortened for better demonstration. Q_i^*, A_i^*: questions and gold answers from the collected dataset, A_i: model predictions.

$\kappa = 0.679$ (substantial agreement).

4 Disagreements between Human and Gold History Evaluation

We now compare the results from the human evaluation and the automatic evaluation with gold history. Note that the two sets of numbers are not directly comparable: (1) the human evaluation reports accuracy, while the automatic evaluation reports F1 scores; (2) the absolute numbers of human evaluation are much higher than those of automatic evaluations. In automatic evaluations, the gold answers cannot capture all possible correct answers to open-ended questions or questions with multiple answers. However, the annotators can evaluate the correctness of answers easily in human evaluations. Nevertheless, we can compare the rankings and the relative gaps between models.

Figure 2 shows different trends between human evaluation and Auto-Gold. Current standard evaluation cannot reflect model performance in human-machine conversations: (1) Human evaluation and Auto-Gold rank BERT and GraphFlow differently; (2) The gap between HAM and ExCorD is significant in Auto-Gold but the two models perform on par in human evaluation.

5 Strategies for Automatic Evaluation

The inconsistency between human evaluation and gold-history evaluation suggests that we need better ways to evaluate and develop our CQA models. When placed in realistic settings, the models never have access to the ground truth (gold answers) and are only exposed to the conversational history and the passage. Intuitively, we can simply replace gold answers by the predicted answers of models and we name this as **predicted-history evaluation** (*Auto-Pred*). Formally, the model makes predictions based on the questions and their own answers:

$$A_i \sim \mathcal{M}(P, Q_1^*, A_1, ..., Q_{i-1}^*, A_{i-1}, Q_i^*).$$

This evaluation has been suggested by several recent works (Mandya et al., 2020; Siblini et al., 2021) which reported a significant performance drop using predicted history. We observe the same performance degradation, shown in Table 1.

However, another issue naturally arises with predicted history: Q_i^*s were written by the dataset annotators based on $(Q_1^*, A_1^*, ..., Q_{i-1}^*, A_{i-1}^*)$ which may become unnatural or invalid if we change the history to $(Q_1^*, A_1, ..., Q_{i-1}^*, A_{i-1})$. We investigate this issue in depth next.

5.1 Predicted History Invalidates Questions

We examined 100 QuAC conversations with the best-performing model (ExCorD) and identified three categories of invalid questions caused by predicted history. We find that 23% of the questions become invalid after using the predicted history. We summarize the types of invalid questions as follows (see Figure 3 for examples):

- **Unresolved coreference** (44.0%). The question becomes invalid for containing either a pronoun

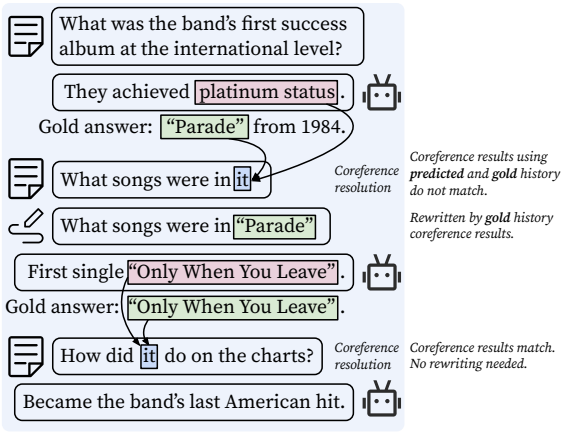


Figure 4: An example of question rewriting. We rewrite the second question with referent in the gold history, because predicted and gold history have different coreference results. We do not rewrite the third question as coreference results are the same.

or a definite noun phrase that refers to an entity unresolvable without the gold history.

- **Incoherence** (39.1%). The question is incoherent with the conversation flow (e.g., mentioning an entity non-existent in predictions). While humans may still answer the question using the passage, this leads to an unnatural conversation and a train-test discrepancy for models.
- **Correct answer changed** (16.9%). The answer to this question with the predicted history changes from when it is based on the gold history.

We further analyze the reasons for the biggest "unresolved coreference" category and find that the model either gives an incorrect answer to the previous question ("incorrect prediction", 17.5%), or the model predicts a different (yet correct) answer to an open question ("open question", 16.3%), or the model returns `CANNOT ANSWER` incorrectly ("no prediction", 4.2%), or the gold answer is longer than prediction and the next question depends on the extra part ("extra gold information", 6.0%).

Invalid questions result in compounding errors as the model is not able to parse them correctly, which may further affect how the model interprets the next questions. Since "unresolved coreference" accounts for most of invalid questions, we aim to address them with a better automatic evaluation.

5.2 Evaluation with Question Rewriting

Among all the invalid question categories, "unresolved coreference" questions are the most critical ones. They lead to incorrect interpretations of questions and hence wrong answers. We propose to improve our evaluation by incorporating a state-of-the-art coreference resolution system to automatically detect invalid questions categorized as "unresolved coreference". More specifically, we use the coreference model from Lee et al. (2018) in `AllenNLP` (Gardner et al., 2018). We make the assumption that if the coreference model resolves mentions in Q_i^* differently using gold history $(Q_1^*, A_1^*, ..., A_{i-1}^*, Q_i^*)$ and predicted history $(Q_1^*, A_1, ..., A_{i-1}, Q_i^*)$, then Q_i^* is identified as having an unresolved coreference issue.

Detecting invalid questions. The inputs to the coreference model for Q_i^* are the following:

$$S_i^* = [BG; Q_{i-k}^*; A_{i-k}^*; Q_{i-k+1}^*; A_{i-k+1}^*; ...; Q_i^*]$$
$$S_i = [BG; Q_{i-k}^*; A_{i-k}; Q_{i-k+1}^*; A_{i-k+1}; ...; Q_i^*],$$

where BG is the background [4], S_i^* and S_i denote the inputs for gold and predicted history. We are only interested in the entities mentioned in the current question Q_t^* and we filter out named entities (e.g., the *National Football League*) because they can be understood without coreference resolution. After the coreference model returns entity cluster information given S_i^* and S_i, we extract a list of entities $E^* = \{e_1^*, ..., e_{|E^*|}^*\}$ and $E = \{e_1, ..., e_{|E|}\}$. We say Q_i^* is *valid* only if $E^* = E$, that is,

$$|E^*| = |E| \text{ and } e_j^* = e_j, \forall e_j \in E,$$

assuming e_j^* and e_j has a shared mention in Q_i^*. We determine whether $e_j^* = e_j$ by checking if $\text{F1}(s_{j,1}^*, s_{j,1}) > 0$, where $s_{j,1}^*$ is the first mention of e_j^* and $s_{j,1}$ is the first mention of e_j, and F1 is the word-level F1 score, i.e. $e_j^* = e_j$ as long as their first mentions have word overlap.

Question rewriting through entity substitution. Our first strategy is to substitute the entity names in Q_i^* with entities in E^*, if Q_i^* is invalid. The rewritten question, instead of the original one, will be used in the conversation history and fed into the model. We denote this evaluation method as **rewritten-question evaluation** (*Auto-Rewrite*), and Figure 4 illustrates a concrete example. Our algorithm rewrites ∼12% of the questions for all of the models. An analysis of rewritten questions' quality is provided in Appendix B.

[4] QuAC provides a short background for each passage, which is the first paragraph of the article the passage is from. It is empirically helpful for associating different spans in the conversation to the same entity mention.

	Overall Performance				Answerable Q. Performance			
	BERT	GraphFlow	HAM	ExCorD	BERT	GraphFlow	HAM	ExCorD
Auto-Gold (F1)	63.2	64.9	65.4	67.7	61.8	66.6	64.5	66.4
Auto-Pred (F1)	54.6	49.6	57.2	61.2	52.7	54.5	54.6	59.2
Auto-Rewrite (F1)	54.5	48.2	57.3	61.9	51.2	51.9	55.1	59.7
Auto-Replace (F1)	54.2	47.8	57.1	61.7	50.7	51.7	54.8	59.7
Human (Accuracy)	82.6	81.0	87.8	87.9	75.9	83.2	84.8	85.3

Table 1: Model performance in automatic and human evaluations. We report *overall performance* on all questions and also performance on *answerable questions* only.

Question replacement using CANARD. Alternative to automatically rewriting questions, we also tried replacing the invalid questions with its human-written context-independent counterpart from CANARD (Elgohary et al., 2019), which we denote as **replaced-question evaluation** (*Auto-Replace*). Since collecting context-independent questions is expensive, Auto-Replace is limited to evaluating models trained with QuAC, thus we do not treat this as a generic method for CQA evaluation.

6 Automatic vs Human Evaluation

In this section, we compare human evaluation results with all the automatic evaluations we have introduced: gold-history evaluation (Auto-Gold), predicted-history evaluation (Auto-Pred), and our proposed Auto-Rewrite and Auto-Replace. We first explain how we compare different evaluation results and then discuss the findings.

6.1 Agreement Metrics

Model performance and rankings. We first consider using model performance reported by different evaluation methods. Considering numbers of automatic and human evaluations are not directly comparable, we also calculate models' rankings and compare whether the rankings are consistent between automatic and human evaluations. Model performance is reported in Table 1. In human evaluation, GraphFlow $<$ BERT $<$ HAM $\approx$ ExCorD; in Auto-Gold, BERT $<$ GraphFlow $<$ HAM $<$ ExCorD; in other automatic evaluations, GraphFlow $<$ BERT $<$ HAM $<$ ExCorD.

Unanswerable statistics. Percentage of unanswerable questions is an important attribute for conversations. Automatic evaluations using static datasets have a fixed number of unanswerable questions, while in human evaluation, the amount of unanswerable questions asked by human annotators varies with different models. The statistics of unanswerable questions is shown in Table 2.

Human Evaluation				QuAC
BERT	GraphFlow	HAM	ExCorD	
34.6	20.6	34.1	33.2	20.2

Table 2: Percentage of unanswerable questions (%) in each model's human evaluation and the original QuAC dataset (used for all automatic evaluations).

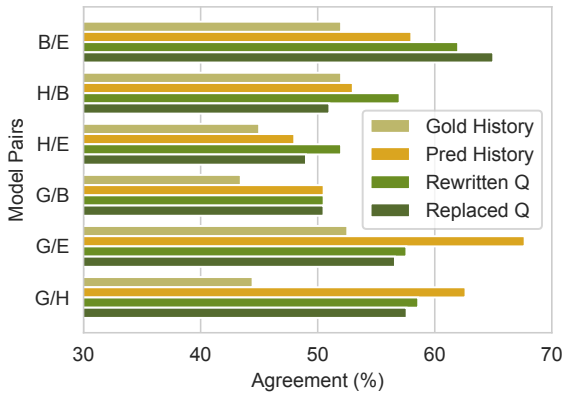


Figure 5: Pairwise agreement of different model pairs comparing automatic evaluations to human evaluation. B: BERT; G: GraphFlow; H: HAM; E: ExCorD.

Pairwise agreement. For a more fine-grained evaluation, we perform a passage-level comparison for every pair of models. More specifically, for every single passage we use one automatic metric to decide whether model A outperforms model B (or vice versa) and examine the percentage of passages that the automatic metric agrees with human evaluation. For example, if the pairwise agreement of BERT/ExCorD between human evaluation and Auto-Gold is 52%, it means that Auto-Gold and human evaluation agree on 52% passages in terms of

	Predicted Unans. Q.				Precision				Recall			
	B	G	H	E	B	G	H	E	B	G	H	E
Auto-Gold	27.1	21.5	27.1	28.3	56.8	62.3	57.1	57.9	68.1	59.3	68.4	72.5
Auto-Pred	27.8	13.8	28.6	28.9	50.0	53.9	52.3	53.3	61.4	33.0	66.1	68.2
Auto-Rewrite	27.3	13.1	25.1	26.0	48.6	55.0	52.4	53.9	65.7	35.7	65.1	69.4
Auto-Replace	27.5	12.9	25.2	25.7	48.6	54.2	52.1	53.8	66.1	34.7	64.9	68.4
Human	42.3	14.7	37.2	36.0	75.0	93.0	86.8	87.4	95.2	72.5	93.7	93.3

Table 3: The percentage of models' predicted unanswerable questions, and the precision and recall for detecting unanswerable questions in different evaluations. B: BERT; G: GraphFlow; H: HAM; E: ExCorD.

which model is better. Higher agreement means the automatic evaluation is closer to human evaluation. Figure 5 shows the results of pairwise agreement.

6.2 Key Findings

Automatic evaluations have a significant distribution shift from human evaluation. We draw this conclusion from the three following points.

- Human evaluation shows a much higher model performance than all automatic evaluations, as shown in Table 1. Two reasons caused this huge discrepancy: (a) Many CQA questions have multiple possible answers, and it's hard for the static dataset in automatic evaluations to capture all the answers. It is not an issue in human evaluation for all answers are judged by human evaluators. (b) There are more unanswerable questions and open questions in human evaluation (reason discussed in the next paragraph), which are relatively easy.
- Human evaluation has a much higher unanswerable question rate, as shown in Table 2. The reason is that in human-human data collection, the answers are usually correct and the questioners can ask followup questions upon the high-quality conversation; in human-machine interactions, since the models can make mistakes, the conversation flow is less fluent and it is harder to have followup questions. Thus, questioners chatting with models tend to ask more open or unanswerable questions. This also suggests that current CQA models are far from perfection.
- All automatic evaluation methods have a pairwise agreement lower than 70% with human evaluation, as demonstrated in Figure 2.

Auto-Rewrite is closer to human evaluation. First, we can clearly see that among all automatic evaluations, Auto-Gold deviates the most from the human evaluation. From Table 1, only Auto-Gold shows different rankings from human evaluation, while Auto-Pred, Auto-Rewrite, and Auto-Replace show consistent rankings to human judgments.

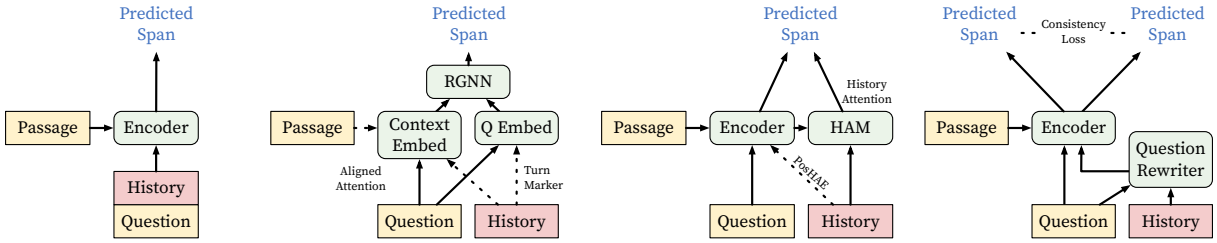
In Figure 2, we see that Auto-Gold has the lowest agreement with human evaluation; among others, Auto-Rewrite better agrees with human evaluation for most model pairs. Surprisingly, Auto-Rewrite is even better than Auto-Replace – which uses human annotated context independent questions – in most cases. It shows that our rewriting policy can better reflect the real-world CQA performance.

7 Towards Better Conversational QA

With insights drawn from human evaluation and comparison with automatic evaluations, we discuss the impact of different modeling strategies, as well as future directions towards better CQA systems.

Modeling question dependencies on conversational context. When we focus on *answerable questions* (Table 1), we notice that GraphFlow, HAM and ExCorD perform much better than BERT. We compare the modeling differences of the four systems in Figure 6, and identify that all the three better systems explicitly model the question dependencies on the conversation history and the passage: both GraphFlow and HAM highlight repeated mentions in questions and conversation history by special embeddings (turn marker and PosHAE) and use attention mechanism to select the most relevant part from the context; ExCorD adopts a question rewriting module that generates context-independent questions given the history and passage. All those designs help models better understand the question in a conversational context.

Unanswerable question detection. Table 3 demonstrates models' performance in detecting *unanswerable questions*. We notice that GraphFlow predicts much fewer unanswerable questions



(a) BERT (b) GraphFlow (c) HAM (d) ExCorD

Figure 6: Modeling structures of BERT, GraphFlow, HAM, and ExCorD.

than the other three models, and has a high precision and a low recall in unanswerable detection. This is because GraphFlow uses a separate network for predicting unanswerable questions, which is harder to calibrate, while the other models jointly predict unanswerable questions and answer spans.

This behavior has two effects: (a) GraphFlow's overall performance is dragged down by its poor unanswerable detection result (Table 1). (b) In human evaluation, annotators ask fewer unanswerable questions with GraphFlow (Table 2) – when the model outputs more, regardless of correctness, the human questioner has a higher chance to ask passage-related followup questions. Both suggest that how well the model detects unanswerable questions significantly affects its performance and the flow in human-machine conversations.

Optimizing towards the new testing protocols. Most existing works on CQA modeling focus on optimizing towards Auto-Gold evaluation. Since Auto-Gold has a large gap from the real world evaluation, more efforts are needed in optimizing towards the human evaluation, or Auto-Rewrite, which better reflects human evaluation. One potential direction is to improve models' robustness given noisy conversation history, which simulates the inaccurate history in real conversations that consists of models' own predictions. In fact, prior works (Mandya et al., 2020; Siblini et al., 2021) that used predicted history in training showed that it benefits the models in predicted-history evaluation.

8 Related Work

Conversational question answering. In recent years, several conversational question answering datasets have emerged, such as QuAC (Choi et al., 2018), CoQA (Reddy et al., 2019), and DoQA (Campos et al., 2020). Different from single-turn QA datasets (Rajpurkar et al., 2016), CQA requires the model to understand the question in the context of conversational history. There have been many methods proposed to improve CQA performance (Ohsugi et al., 2019; Chen et al., 2020; Qu et al., 2019; Kim et al., 2021) and significant improvement has been made on CQA benchmarks.

Besides text-based CQA tasks, there also exist CQA benchmarks that require other forms of modeling ability, such as combining textual evidence with background knowledge (Saeidi et al., 2018), utilizing structured knowledge base (Saha et al., 2018; Guo et al., 2018), as well as CQA in other modalities (Das et al., 2017).

Evaluation with predicted history. Only recently has it been noticed that the current method of evaluating CQA models is flawed. Mandya et al. (2020); Siblini et al. (2021) point out that using gold answers in history is not consistent with the real-world scenario and propose to use predicted history for evaluation. Different from prior work, in this paper, we conduct a large scale human evaluation to support our claims, identify the issues with predicted history, and propose rewriting questions to further mitigate the gap to human evaluation.

9 Conclusion

In this work, we carry out the first large-scale human evaluation on CQA systems. We show that current standard automatic evaluation with gold history cannot reflect models' performance in human evaluation, and that human-machine conversations have a large distribution shift from static CQA datasets of human-human conversations. To tackle these problems, we propose to use predicted history with rewriting invalid questions for evaluation, which reduces the gap between automatic evaluations and the real-world human evaluation. We also use the human evaluation results to analyze current CQA systems and identify promising directions for future development.

References

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Association for Computational Linguistics (ACL)*, pages 7302–7314.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2020. Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2174–2184.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-action: Conversational question answering over a large-scale knowledge base. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2942–2951.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In *Association for Computational Linguistics (ACL)*, pages 6130–6141.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 687–692.

Angrosh Mandya, James O' Neill, Danushka Bollegala, and Frans Coenen. 2020. Do not let the history haunt you: Mitigating compounding errors in conversational question answering. In *International Conference on Language Resources and Evaluation (LREC)*, pages 2017–2025.

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 *Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 79–84.

Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 11–17.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019. Attentive history selection for conversational question answering. In *ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1391–1400.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association of Computational Linguistics (TACL)*, pages 249–266.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2087–2097.

Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. *arXiv preprint arXiv:1801.10314*.

Wissam Siblini, Baris Sayil, and Yacine Kessaci. 2021. Towards a more robust evaluation for conversational question answering. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1028–1034.

A Human Evaluation Statistics

Table 4 shows the human evaluation statistics, including numbers of conversations and questions regarding each model.

	BERT	GraphFlow	HAM	ExCorD	QuAC Dev
#Conv	357	359	373	357	1,000
#Q	3,755	3,666	3,959	3,679	7,354

Table 4: Numbers of conversations and questions collected in human evaluation, using 100 QuAC development set passages. We also put the statistics for QuAC development set for reference.

B Quality of Rewriting Questions

To analyze how well Auto-Rewrite does in detecting and rewriting questions, we manually check 100 conversations of ExCorD from the QuAC development set. We find that Auto-Rewrite can detect invalid questions with a precision of 72% and a recall of 72%. We notice that the coreference model sometimes detects the pronoun of the main character in the passage, which almost shows up in every question, as insolvable. This issue causes the low precision but is not a serious problem in our case – whether rewriting the pronoun of the main character does not affect models' prediction much, because the model always sees the passage and knows who the main character is.

Among all correctly detected invalid questions, we further check the quality of rewriting, and in 68% of the times Auto-Rewrite gives a correct context-independent questions. The most common error is being ungrammatical: For example, using the gold history of "... Dee Dee claimed that Spector once ***pulled*** a gun on him", the original question "Did they arrest him for doing *this*?" was rewritten to "Did they arrest Phillip Harvey Spector for doing ***pulled***?" While this creates a distribution shift on question formats, it is still better than putting an invalid question in the flow.

C Importance of Explicit Dependency Modeling.

Figure 7 gives an example where GraphFlow, HAM and ExCorD correctly resolve the question phrase from long conversation history while BERT failed. This is caused by BERT's lack of explicit question dependency modeling.

Tom McCall – Vortex I
... McCall decided to hold a rock festival at Milo McIver State Park, Oregon called "Vortex I: A Biodegradable Festival of Life"...

Q_1^*:	Was Vortex I popular?	
B:	*The festival*, "The Governor's Pot Party" ...	✓
G/H/E:	*The festival*, "The Governor's Pot Party" ...	✓
...		
Q_4:	Who played at *the festival*?	
B:	`CANNOT ANSWER`	✗
G/H/E:	Gold, The Portland Zoo, Osceola, Fox...	✓

Figure 7: An example of BERT failing to resolve *the festival* in Q_4^*, while all other models with explicit dependency modelings succeeded.