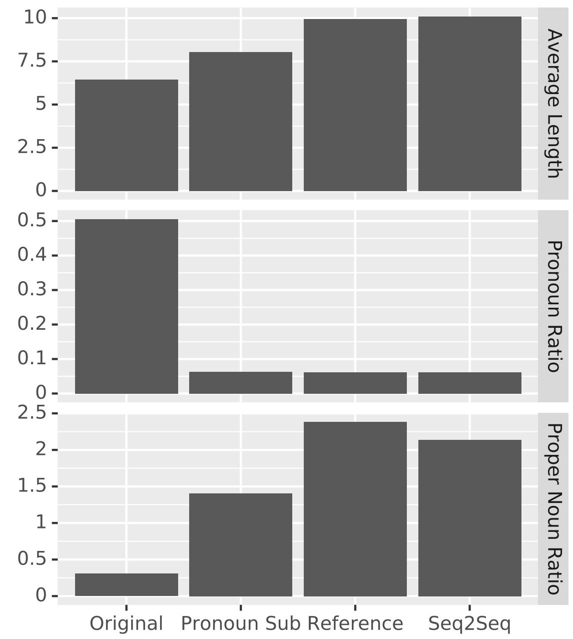

Figure 2: Human rewrites are longer, have fewer pronouns, and have more proper nouns than the original QuAC questions. Rewrites are longer and contain more proper nouns than our Pronoun Sub baseline and trained Seq2Seq model.

neural sequence-to-sequence improves 2–4 BLEU points over naive baselines, it is still 9 BLEU points below human-accuracy. We analyze sources of errors in the following section.

5 Dataset and Model Analysis

We analyze our dataset with automatic metrics after validating the reliability of our data (Section 3). We compare our dataset to the original QuAC questions and to automatically generated questions by our models. Then, we manually inspect the sources of rewriting errors in the seq2seq baseline.

5.1 Anaphora Resolution and Coreference

Our rewrites are longer, contain more nouns and less pronouns, and have more word types than the original data. Machine output lies in between the two human-generated corpora, but quality is difficult to assess. Figure 2 shows these statistics. We motivate our rewrites by exploring linguistic properties of our data. Anaphora resolution and coreference are two core NLP tasks applicable to this dataset, in addition to the downstream tasks evaluated in Section 4.

Pronouns occur in 53.9% of QuAC questions. Questions with pronouns are more likely to be ambiguous than those without any. Only 0.9% of these have pronouns that span more than one category (e.g., ‘she’ and ‘his’). Hence, pronouns within a single sentence are likely unambiguous. However, 75.0% of the aggregate history has pronouns and the percentage of mixed category pronouns increase to 27.8% of our data. Therefore, pronoun disambiguation potentially becomes a problem for a quarter of the original data. An example is provided in Table 4.

Label	**Text**
QUESTION	How long did he stay there?
REWRITE	How long did Cito Gaston stay at the Jays?
HISTORY	*Cito Gaston* **Q:** What did Gaston do after the world series? ... **Q:** Where did he go in 2001? **A:** In 2002, he was hired by the Jays as special assistant to president and chief executive officer Paul Godfrey.

Table 4: An example that had over ten flagged proper nouns in the history. Rewriting requires resolving challenging coreferences.

Approximately one-third of the questions generated by our pronoun-replacement baseline are within 85% string similarity to our rewritten questions. That leaves two-thirds of our data that cannot be solved with pronoun resolution alone.

5.2 Model Analysis

By manually examining the predictions of the seq2seq model, we notice that the main source of errors is that the model tends to find a short path to completing the rewrites. That often results in *under-specified questions* as in Example 1 in Table 5, *question meaning change* as in Example 2 or *meaningless questions* as in Example 3.

Another source of errors is having related entities mentioned in the context as Example 4 in Table 5, where the model confused “Copa America” with “Argentina”. The model also struggles with listing multiple entities mentioned in different parts of the context. Example 5 in Table 5 show the output and the reference rewrites of the question *“Did she have any more works than those 3?”*, where two of the three entities—“United States of Banana”, “La Comedia” and “Asalto al tiempo”—are lost in the rewrite.

	Seq2Seq output	**Reference**
1	What did Chamberlain's men do?	What did Chamberlain's men do during the Battle of Gettysburg?
2	How many games did Ozzie Smith win?	How many games did the Cardinals win while Ozzie Smith played?
3	Did 108th get to the finals?	Did the US Women's Soccer Team get to the finals in the 1999 World Cup?
4	Did Gabriel Batistuta reside in any other countries, besides touring in the Copa America?	Besides Argentina, did Gabriel Batistuta reside in any other countries?
5	Did La Comedia have any more works than La Comedia 3?	Did Giannina Braschi have any more works than United States of Banana, La Comedia and Asalto al tiempo?

Table 5: Example erroneous rewrites generated by the Seq2Seq models and their corresponding reference rewrites. The dominant source of error is the model tendency to produce short rewrites (Examples 1–3). Related entities (Copa America and Argentina in Example 4) distract the model. The model struggles with listing multiple entities mentioned in different parts of the context (Example 5).

6 Related Work and Discussion

Recent work in CQA has used simple concatenation (Elgohary et al., 2018), sequential neural models (Huang et al., 2019), and transformers (Qu et al., 2019a) for modeling the interaction between the conversation history, the question and reference documents. Some of the components in those models, such as relevant history turn selection (Qu et al., 2019b), can be adopted in question rewriting models for our task. An interesting avenue for future work is to incorporate deeper context, either from other modalities (Das et al., 2017) or from other dialog comprehension tasks (Sun et al., 2019).

Parallel to our work, Rastogi et al. (2019) and Su et al. (2019) introduce utterance rewriting datasets for dialog state tracking. Rastogi et al. (2019) covers a narrow set of domains and the rewrites of Su et al. (2019) are based on Chinese dialog with two-turn fixed histories. In contrast, CANARD has histories of variable turn lengths, covers wider topics, and is based on CQA.

Training question rewriting using reinforcement learning with the task accuracy as reward signal is explored in retrieval-based QA (Liu et al., 2019) and in MRC (Buck et al., 2018). A natural question is whether reinforcement learning could learn to retain the necessary context to rewrite questions in CQA. However, our dataset could be used to pretrain a question rewriter that can further be refined using reinforcement learning.

More broadly, we hope CANARD can drive human-computer collaboration in QA (Feng and Boyd-Graber, 2019). While questions typically vary in difficulty (Sugawara et al., 2018), existing research either introduces new benchmarks of difficult (adversarial) stand-alone questions (Dua et al., 2019; Wallace et al., 2019, inter alia), or models that simplify hard questions through paraphrasing (Dong et al., 2017) or decomposition (Talmor and Berant, 2018). We aim at studying QA models that can ask for human assistance (feedback) when they struggle to answer a question.

The reading comprehension setup of CQA provides a controlled environment where the main source of difficulty is interpreting a question in its context. The interactive component of CQA also provides a natural mechanism for improving rewriting. When the computer cannot understand (rewrite) a question because of complicated context, missing world knowledge, or upstream errors (Peskov et al., 2019) in the course of a conversation, it should be able to ask its interlocutor, "can you unpack that?" This dataset helps start that conversation; the next steps are developing and evaluating models that efficiently decide when to ask for human assistance, and how to best use this assistance.

Acknowledgments

We thank Jo Shoemaker and Michelle Yuan for crowdsourcing feedback. This work was supported by NSF grant IIS-1822494. Boyd-Graber and Peskov were also supported by IIS-1748663. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. 2018. Ask the right questions: Active question reformulation with reinforcement learning. In *Proceedings of the International Conference on Learning Representations*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of Empirical Methods in Natural Language Processing*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Computer Vision and Pattern Recognition*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. Dataset and baselines for sequential open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Shi Feng and Jordan Boyd-Graber. 2019. What AI can do for me: Evaluating machine learning interpretations in cooperative play. In *International Conference on Intelligent User Interfaces*.

Hsin-Yuan Huang, Eunsol Choi, and Wen tau Yih. 2019. FlowQA: Grasping flow in history for conversational machine comprehension. In *Proceedings of the International Conference on Learning Representations*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of Association for Machine Translation in the Americas*.

Ye Liu, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S Yu. 2019. Generative question refinement with deep reinforcement learning in retrieval-based QA system. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Denis Peskov, Joe Barrow, Pedro Rodriguez, Graham Neubig, and Jordan Boyd-Graber. 2019. Mitigating noisy inputs for question answering. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. BERT with history modeling for conversational question answering. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, and Mohit Iyyer. 2019b. Attentive history selection for conversational question answering. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Pushpendre Rastogi, Arpit Gupta, Tongfei Chen, and Lambert Mathias. 2019. Scaling multi-domain dialogue state tracking via query reformulation. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the Association for Computational Linguistics*.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the Conference of the European Chapter of the Association for Computational Linguistics*.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance