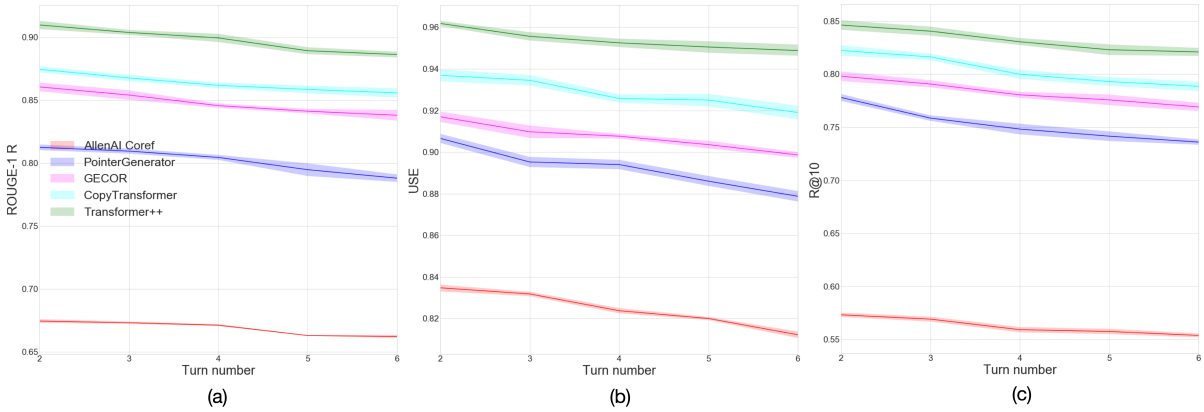

Figure 3: Rouge-1R, USE and R@10 metrics of baseline co-reference model, top-3 encoder-decoder models, and Transformer++ model based on dialogue turn number

PointerGenerator uses a bi-LSTM encoder and a pointer-generator decoder, which allows to copy and generate tokens (Elgohary et al., 2019).

GECOR uses two bi-GRU encoders, one for user utterance and other for dialogue context, and a pointer-generator decoder previously proposed for task-oriented dialogues (Quan et al., 2019).

Generator is a Transformer decoder model with a language modeling head (linear layer in the size of the vocabulary) (Radford et al., 2019).

Generator + Multiple-choice model has a second head for the auxiliary classification task that distinguishes between the correct rewrite and several noisy rewrites as negative samples (inspired by TransferTransfo (Wolf et al., 2019b)).

CopyTransformer uses one of the attention heads of the Transformer as a pointer to copy tokens from the input sequence directly (Gehrmann et al., 2018).

Transformer++ model has two language modeling heads that produce separate vocabulary distributions, which are then combined via a parameterized weighted sum (the coefficients are produced by combining the output of the first attention head and the input embeddings).

7.2 BERTserini

We implemented BERTserini following Yang et al. (2019) We use the standard BM25 ranking for passage retrieval with $k_1 = 0.82$, $b = 0.68$, which was previously found to work well for passage retrieval on MS MARCO. We then retrieve the top-100 relevant passages per question. Afterwards, we use BERT-Large fine-tuned for the task of reading comprehension. This model takes a question and each of the relevant passages as input and produces the answer span (Wolf et al., 2019a). BERT-Large produces a score (S_{BERT}), which is combined with the retrieval score for each of the passages (S_{Anserini}) through simple linear interpolation:

$$S = (1 - \mu) \cdot S_{\text{Anserini}} + \mu \cdot S_{\text{BERT}}$$

We pick the span with the highest score S as the answer. The parameter $\mu \in [0, 1]$ was tuned using a 10% random subset of the QReCC training set withheld from the BERT-Large training (we found $\mu = 0.7$ to work best).

BERT-Large was trained on human rewrites from the QReCC training set, and evaluated on the test set using either the original questions, human rewrites or the rewrites produced by Transformer++. The model is trained to either predict an answer span or predict that the passage does not contain an answer. "No answer" for the question is predicted only when neither of the relevant passages predicts an answer span. The model was trained on 480K paragraphs that contain the correct answers and 5K of other paragraphs as negative samples (see Appendix A.3 for more details).

8 Baseline Results

We use the results of QR to select the best model and then use it for the end-to-end QA task. Question rewrites are used as input for both passage retrieval and reading comprehension tasks. The effectiveness of the QR component is compared with the end-to-end model conditioned on the conversational context.

Table 5: Mean reciprocal rank, recall@10, and recall@100 for passage retrieval on test set questions.

Rewrite Type	MRR	R@10	R@100
Original	0.0343	6.12	11.71
Transformer++	0.1586	26.52	41.51
Human	0.1994	32.78	49.36

8.1 Question Rewriting Effectiveness

We analyze the effectiveness of our QR models by doing a 5-fold cross validation and obtaining the best performing metrics. Figure 3 contains 3 plots showing ROUGE 1-R, USE and R@10 across 5 turns. We start with the second turn because the first turn always is a self-contained query. The metrics across turns also stay stable with the same result for all the models. The Transformer++ model is stable with little variance in terms of its maximum and minimum metric values across all the best performing metrics.

Our evaluation results are summarized in Table 4. All generative models outperform the state-of-the-art coreference resolution model (AllenAI Coref). We noticed that PointerGenerator which employs a bi-LSTM encoder with a copy and generate mechanism outperforms Generator using Transformer alone. We could not find evidence that pretraining with an auxiliary regression task can improve the QR model effectiveness (Generator + Multiple-choice). Use of two separate bi-GRU encoders for the query and conversation context further improved the QR effectiveness (GECOR). Modeling both copying and generating the tokens from the input sequence employing the Transformer helped improve the effectiveness of the QR model (Copy-Transformer) compared to other existing generative models. Finally, obtaining the final distribution by computing token probabilities and weighting question and context vocabulary distributions with those probabilities helped improve over the best performing generative model (Transformer++).

8.2 Question Answering Effectiveness

Table 5 shows the mean reciprocal rank (MRR), R@10, and R@100 of using the original, Transformer++, and human rewritten questions. R@k is averaged across all questions. For a question, if R@k is 1.0, it means that there is a passage in the top-k at any rank such that the passage is relevant; and 0.0 otherwise. Table 6 shows the standard F1 and Exact Match metrics for extractive QA for each type of input question. In the "End-to-End" setting, the retrieval score was combined with the BERT reader score to determine the final span. In the "Known Context" setting, we use the relevant passage from the web page indicated by the human annotator, i.e., without passage retrieval. In the "Extractive Upper Bound" setting, we use a heuristic to find the answer span with the highest F1 score among the top-100 retrieved passages with human rewrite. This setup indicates the best the reader can do given the retrieval results.

Table 6: Mean F1 and Exact Match scores (%) on passages for extractive QA. "Known Context" assumes perfect retrieval. The "Extractive Upper Bound" assumes perfect single document span extraction.

Setting	Rewrite Type	F1	EM
End-to-End	Original	9.07	0.32
	Transformer++	19.10	1.01
	Human	21.82	1.23
Known Context	Original	17.24	1.90
	Transformer++	32.34	4.04
	Human	36.42	4.70
Extractive Upper Bound		75.45	25.07

The upper bound on the answer span extraction (F1 = 75.45) highlights the need for more sophisticated QA techniques than the standard reading comprehension approaches can offer now. Some answer texts in QReCC were paraphrased or summarised using multiple passages from the same web page. Abstractive approaches to answer generation are necessary to close this gap.

Even using single document span extraction techniques, there is a large room for improvement. Comparing "Known Context" to "End-to-End" we see losses introduced by the retrieval step, and comparing the "Extractive Upper Bound" to "Known Context" we see the sizeable margin of improvement available even for extractive models. This shows that even with competitive baselines the QA tasks are all far from solved.

In both Table 5 and 6 we see that human rewritten questions more than double the effectiveness of using original questions. In the absence of human rewritten questions, using Transfomer++ elevates the effectiveness of the QA tasks, getting it much closer to that proffered by human-level QR.

9 Conclusion

We introduced the QReCC dataset for open-domain conversational QA. QReCC is the first dataset to

cover all the subtasks relevant for conversational QA, which include question rewriting, passage retrieval and reading comprehension. We also set the first end-to-end baseline results for QReCC by evaluating an open-domain QA model in combination with a QR model. We presented a systematic comparison of existing automatic evaluation metrics on assessing the quality of question rewrites and show the metrics that best proxy human judgement. Our empirical evaluation shows that QR provides an effective solution for resolving both ellipsis and co-reference that allows to use existing non-conversational QA models in a conversational dialogue setting. Our end-to-end baselines achieve an F1 score of 19.10, well beneath the 75.45 extractive upper bound, suggesting not only room for improvement in extractive conversational QA, but that more sophisticated abstractive techniques are required to successfully solve QReCC.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset. In *Advances in Neural Information Processing Systems, 2016, NIPS 2016 3-8 December, Barcelona, Spain*.

Bill Byrne, Karthik Krishnamoorthi, Sankar Chinnadhurai, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kim Kyu-Young. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, page 4516–4525.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 169–174.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. Cast 2019: The conversational assistance track overview. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC*, pages 13–15.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 376–380.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5920–5926.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 687–692.