Table 2: Summary statistics for the QReCC dataset.

QReCC	Train	Dev.	Test	All
# questions (Qs)	50.8K	12.7K	16.4K	80.0K
# dialogues	8.7K	2.2K	2.8K	13.6K
max Qs/dialogue	12	12	12	12
avg Qs/dialogue	6	6	6	6
min Qs/dialogue	5	5	5	5
% replacement	53	52	53	52
% insertion	35	36	37	38
% copy	11	11	9	9
% removal	1	1	1	1

setup helps to obtain feedback on the quality of QR with respect to the effectiveness of answer retrieval (see Section 6 for more details on using search results for the evaluating QR). Finally, the question-answer pair is annotated with the link to the web page that was used to produce the answer.

Thereby, every dialogue was produced by the same annotator including the questions, answers and rewrites. This design decision is called *self-dialog technique* that was shown to help improve quality of the data by avoiding some of the challenges observed in simulated dialogues produced by pairs of annotators (Byrne et al., 2019).

A team of 30 professional annotators with a project lead were employed to perform the task. The annotation task was described in the guidelines (see Appendix B for more details). To ensure the quality of the annotations we followed a post-hoc evaluation procedure, in which 5 reviewers go through the dataset and update incorrect examples they identify with consensus.

4 Dialogue Analysis

QReCC contains 13,598 dialogues with 79,952 questions in total. 9.3K dialogues are based on the questions from QuAC; 80 are from TREC CAsT; and 4.4K are from NQ. 9% of questions in QReCC do not have answers. We still retained the question rewrites even if no answer was found on the web. 112 questions were annotated with links to web pages without answer texts, e.g. “May I have a link to road signs in Singapore?”

We prepared three standard dataset splits and ensured that they are balanced in terms of the standard dialogue statistics and the types of QR (see Table 2). We distinguish four types of QR. They differ with respect to the intervention required to resolve contextual dependencies in dialogue. These types can be automatically identified by measuring the difference between an original question Q and a question rewrite R that are represented as sets using the bag-of-words:

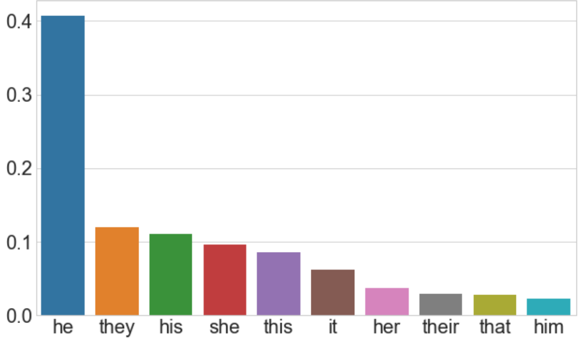

Figure 2: The 10 most frequently replaced tokens in QReCC.

- *Insertion* – new tokens are added to the original question to produce the rewrite (e.g., “What are some of the main types” → “What are some of the main types *of Yoga*?”): $Q \setminus R = \varnothing \land R \setminus Q \neq \varnothing$
- *Removal* – some tokens are removed from the question to produce the rewrite (e.g., “Can you tell me about the C++ language mentioned” → “Can you tell me about the C++ language”): $Q \setminus R \neq \varnothing \land R \setminus Q = \varnothing$
- *Replacement* – some tokens are added and some are removed to produce the rewrite (e.g., “Does it help in reducing stress” → “Does *Yoga* help in reducing stress”): $Q \setminus R \neq \varnothing \land R \setminus Q \neq \varnothing$
- *Copy* – no modification is needed, i.e., the original question is already contextually independent (e.g., “What are common poses in Kundalini Yoga?”): $Q \setminus R = \varnothing \land R \setminus Q = \varnothing$, i.e., $Q = R$

The majority of questions in QReCC (52%) require *Replacement*. Figure 2 shows the tokens that are most frequently replaced in QR. All of them are pronouns that require anaphora resolution. By specifically targeting more rare types of question rewriting in our data collection task we managed to increase the proportion of the *Insertion* cases in our dataset. This allows us to train and evaluate the ability of the model to reconstruct missing context, which cannot be achieved using traditional co-reference resolution approaches.

5 Document Collection

We download the web pages using the answer provenance links provided by the annotators from the Internet Archive Wayback Machine.[2] Then, we complement the relevant pages with randomly sampled web pages that constitute 1% of the Common Crawl dataset identified as English pages. The final collection consists of approximately 14K pages from the Wayback Machine and 9.9M random web pages from the Common Crawl dataset. The scripts for reproducing the passage collection are on GitHub. See Appendix A.2 for more details.

After downloading the pages we extract the textual content from the HTML and split texts into passages of least 220 tokens. After segmentation, we have a total of 54M passages which we index using Anserini (Yang et al., 2017).

We search the passage collection using the human annotated answers to augment the dataset with alternative sources of correct answers. For each document returned, we identify the span in the document that has the highest token overlap (F1) with the human answer. We consider all documents with F1 ≥ 0.8 as relevant. Verifying adequacy of this simple heuristic by human annotators is left for future work.

6 Question Rewriting Metrics Validation

BLEU has typically been used in previous work for measuring the quality of QR (Elgohary et al., 2019; Lin et al., 2020). We conduct a systematic evaluation and compare BLEU with alternative metrics, previously applied in summarization and translation, to ensure the most reliable metrics we can obtain for the model selection. Our evaluation shows that BLEU does not compare favourably with other metrics in evaluating the quality of QR.

Task. We took a random sample of 10K questions and used a seq-to-seq model (Nallapati et al., 2016) trained with questions and conversation context from the QReCC dataset to generate question rewrites. These generated rewrites were compared to the ground truth rewrites produced by human annotators. Different annotators graded each model-generated rewrite with a binary label: 0 (incorrect rewrite) or 1 (correct rewrite). For a question rewrite to be correct it does not have to exactly match the ground truth rewrite, but it should correctly capture the conversational context and be a self-contained question. For example, the model-generated rewrite "What are the global warming dangers?" is a correct rewrite with the ground truth rewrite being "What are the dangers of global warming?". In addition, we also assess the variance of the human assessments. The Pearson correlation between any two annotators on average is 0.94. We observed the mean and the variance to be 0.083 and 0.076 respectively. Performing a two-tail statistical significance test shows the P-value to be 0.0201.

We use several automated metrics to compare the rewrites with the ground truth and compute their Pearson correlation with the human judgements (see Table 3 for results).

Exact Match is a binary variable that indicates the token set overlap applied after the standard preprocessing: lower-casing, stemming, punctuation and stopword removal.

ROUGE (Lin, 2004) reflects similarity between two texts in terms of n-gram overlap (R-1 for unigrams; R-2 for bigrams and R-L for the longest common n-gram). We report the mean for precision (P), recall (R) and F-measure (F).

METEOR (Denkowski and Lavie, 2014) is a machine translation metric based on exact, stem, synonym, and paraphrase matches between words and phrases.

BLEU (Papineni et al., 2002) is a text similarity metric that uses a modified form of precision and n-grams from candidate and reference texts.

Embeddings group several unsupervised approaches that produce a sentence-level vector representation: Universal Sentence Encoder (Cer et al., 2018) and InferSent (Conneau et al., 2017).

Search Results – we use both question rewrites in Google Search and compare the overlap between the produced page ranks in terms of the standard IR metrics: Recall@k for the top-k links, Average Recall (AR) and Normalized Discounted Cumulative Gain (NDCG).

The best performing metric in our experiments (i.e., closest to the human judgement) is the set

[2] We use the version of a web page, which is the closest to the end date of the dialogue collection (November 24, 2019).

Table 3: Comparison of different evaluation metrics in terms of Pearson correlation with the human judgment of the question rewriting quality.

Metrics		Pearson	Metrics	Pearson
Exact Match		0.56	ROUGE-1 P	0.51
Embeddings	**USE**	**0.67**	**ROUGE-1 R**	**0.63**
	InferSent	0.48	ROUGE-1 F	0.61
Search Results	R@1	0.66	ROUGE-2 P	0.54
	R@2	0.72	ROUGE-2 R	0.57
	R@3	0.73	ROUGE-2 F	0.57
	R@4	0.74	ROUGE-L P	0.50
	R@5	0.77	ROUGE-L R	0.61
	R@10	**0.80**	ROUGE-L F	0.58
	AR	0.79	METEOR	0.59
	NDCG	0.74	BLEU	0.58

Table 4: Evaluation results of QR models (mean with 95% confidence intervals). *Human QR metrics are computed across 5 different random samples of 1000 question rewrites from the intersection of QReCC and CANARD conversations.

Model/Metrics	ROUGE-1 R	USE	R@10
AllenAI Coref (Lee et al., 2018)	67.1% ± 10E-4%	82.3% ± 10E-3%	56.1% ± 10E-4%
Generator (Radford et al., 2019)	73.4% ± 0.6%	86.2% ± 0.9%	69.1% ± 0.2%
Generator + Multiple-choice (Wolf et al., 2019b)	74.1% ± 0.5%	86.3% ± 0.4%	70.2% ± 0.1%
PointerGenerator (Elgohary et al., 2019)	80.2% ± 0.8%	89.1% ± 1.1%	75.3% ± 0.3%
GECOR (Quan et al., 2019)	84.1% ± 0.3%	91.8% ± 0.2%	78.1% ± 0.2%
CopyTransformer (Gehrmann et al., 2018)	86.1% ± 0.5%	92.8% ± 0.3%	79.4% ± 0.3%
Transformer++	**89.5% ± 0.4%**	**95.2% ± 0.2%**	**83.2% ± 0.3%**
Human*	94.6% ± 0.2%	97.3% ± 0.1%	87.2% ± 0.1%

overlap of the web search results (**R@10**). The best metrics independent of QA are Universal Sentence Embedding (**USE**) and unigram recall (**ROUGE-1 R**). We provide more details of the metrics performance illustrated with examples and the discussion in Appendix C. We use the set of all three best evaluation metrics to select the optimal QR model for our baseline approach.

7 Baseline Approach

We extend BERTserini (Yang et al., 2019), an efficient approach to open-domain QA, with a QR model to incorporate conversational context. This approach consists of three stages: (1) QR, (2) PR and (3) RC. First, a model is trained to generate a stand-alone question given a follow-up question and the preceding question-answer pairs. In the second stage, PR, the top-k relevant passages are retrieved from the index using BM25 using the rewritten question. Finally, in RC, a model is trained to extract an answer span from a passage or predict if the passage is irrelevant. The scores obtained from PR and RC are then combined as a weighted sum to produce the final score. The span with the highest score is chosen as the final answer.

7.1 Question Rewriting

We evaluate a co-reference model and several generative models on the QR subtask using the question rewrites in QReCC and the set of QR metrics selected in Section 6. The best performing model is then used in a combination with BERTserini to set the baseline results for the end-to-end QA task. All our Transformer-based models were initialized with the pretrained weights of GPT-2 (English medium-size) (Radford et al., 2019) and further fine-tuned on question rewrites from the QReCC training set (see Appendix A.1).

AllenAI Coref is the state-of-the-art model for coreference resolution task (Lee et al., 2018). We adapt it for QR with a heuristic that substitutes all coreference mentions with the corresponding antecedents from the cluster.