Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 119–126
Marseille, 20-25 June 2022

CoQAR: Question Rewriting on CoQA

Quentin Brabant, Gwénolé Lecorvé, Lina M. Rojas-Barahona
Orange Innovation
2 Avenue Pierre Marzin. Lannion. France.
{quentin.brabant, gwenole.lecorve, linamaria.rojasbarahona}@orange.com

Abstract

Questions asked by humans during a conversation often contain contextual dependencies, i.e., explicit or implicit references to previous dialogue turns. These dependencies take the form of coreferences (e.g., via pronoun use) or ellipses, and can make the understanding difficult for automated systems. One way to facilitate the understanding and subsequent treatments of a question is to rewrite it into an out-of-context form, i.e., a form that can be understood without the conversational context. We propose CoQAR, a corpus containing 4.5K conversations from the Conversational Question-Answering dataset CoQA, for a total of 53K follow-up question-answer pairs. Each original question was manually annotated with at least 2 at most 3 out-of-context rewritings. CoQAR can be used in the supervised learning of three tasks: question paraphrasing, question rewriting and conversational question answering. In order to assess the quality of CoQAR's rewritings, we conduct several experiments consisting in training and evaluating models for these three tasks. Our results support the idea that question rewriting can be used as a preprocessing step for question answering models, thereby increasing their performances.

Keywords: question rewriting, conversational question answering, question paraphrasing

1. Introduction

Conversational Question Answering (CQA) (Reddy et al., 2019; Choi et al., 2018; Saha et al., 2018) is a task in which a system interacts with a so-called *student*. The interaction takes the form of a conversation, where the student asks questions, and the system is expected to provide the right answers. In this paper we focus on the case where the system searches for answers in a text passage, although settings relying on structured data (e.g. knowledge bases) also exist (Saha et al., 2018). Compared to non-conversational question answering (or QA for short), the system faces an additional difficulty: each question is asked in a *conversational context* that consists in previous conversation turns; implicit references to the conversational context may happen in the form of ellipses and coreferences, making the understanding of questions more difficult for the system.

One way to overcome this difficulty is Question Rewriting (QR), which consists in rewriting each original (*in-context*) question into an *out-of-context* question that is understandable by itself, i.e., that can be answered without knowing the conversational context. Vakulenko et al. (2021) argue in favor of this approach by experimentally showing that adding QR as a preprocessing step of CQA models can improve their performances. They also claim that QR models offer several advantages, including the possibility of *reuse*: a same QR model can be used as a preprocessing step for several existing (conversational or non-conversational) QA models and datasets. In particular, any existing non-conversational QA model (see, e.g., (Rajpurkar et al., 2018; Usbeck et al., 2018)) can be immediately used for CQA.

In this paper, we present the CoQAR corpus, which is an annotated subset of the CQA corpus CoQA (Reddy et al., 2019). CoQAR was obtained by asking specialised native speakers to annotate original questions with at least two and at most three distinct *out-of-context* rewritings. Our contribution is two-fold.

Firstly, we provide CoQAR, which contains high-quality questions rewritings. The corpus is publicly available[1]; moreover, its annotations were conducted in accordance to ethical concerns: every annotator involved was properly hired.

Secondly, we assess the quality of the annotations of CoQAR through several experiments. We train Question Rewriting (QR) models, as well as Question Paraphrasing (QP) models on CoQAR and other datasets. We then rate these models' outputs via human evaluation. We also evaluate QR models as preprocessing steps of (conversational and non-conversational) QA models. To this end, we compare the performance of a stat-of-the-art QA model with and without QR.

Our results support the claim of Vakulenko et al. (2021) that QR models can be successfully used in combination with existing QA models. Indeed, we found that adding QR as a preprocessing step boosts the performances of QA models and allows reusing non-conversational state-of-the-art QA systems while reducing performance degradation on CoQA.

In the remainder of this paper we present the related work in Section 2. We introduce CoQAR in Section 3. We talk about the NLP task we use to evaluate the proposed annotations in Section 4. The evaluation and discussion are presented in Section 5 and Section 6, respectively.

[1]The COQAR dataset is publicly available at `https://github.com/Orange-OpenSource/COQAR`

passage	This is the story of a young girl and her dog. The young girl and her dog set out a trip into the woods one day. Upon entering the woods the girl and her dog found that the woods were dark and cold [...].
question	What is the story about?
rewrittings	What is the subject of this story? Who are the two main characters in the story? Who is this story centered on?
answer	A girl and a dog.
answer span	This is the story of a young girl and her dog.
question	What were they doing?
rewrittings	What were the girl and her dog up to? What did the girl and her dog decide to do? What was the activity of the girl and the dog for the day?
answer	Set on on a trip
answer span	The young girl and her dog set out a trip
question	where?
rewrittings	Where did the girl and her dog go on a trip? What location did the girl and her dog journey to? What place did the girl and her dog go on that day?
answer	the woods
answer span	set out a trip into the woods

Table 1: Beginning of a passage from CoQAR and of the corresponding conversation.

2. Related Work

CoQA (Reddy et al., 2019) is a Conversational Question Answering dataset that was originally created for measuring the ability of machines to handle conversational question answering. It contains 8k conversations, which sum up to 127k questions with answers. Each dialogue was produced by two human annotators, one *student* asking questions, and one *teacher* providing answers. Each conversation is about a piece of text called *passage*. The questions are conversational, while each answer is provided in two forms: (1) the answer per-say, which is a short piece of text (not necessarily a full sentence); (2) the *answer span*, which is a quote from the passage from which the answer is deduced. Many answers are a subsequence of the answer span; however, this is not always the case. For example, the answer to a yes/no questions is "yes" or "no", although those word usually do not appear in the answer span. Each passage belongs to one of seven domains; two of these domains only appear in the test set. Many questions require pragmatic reasoning, which makes CoQA a challenging evaluation dataset for conversational question answering systems. Moreover, the authors estimate that 70% of the questions cannot be correctly understood without taking into account the context established during previous dialogue turns. Finally, some of those questions are not answerable based on the passage. The right answer to these question is represented by the special "unknown" string.

Similar to our work, the corpus CANARD (Elgohary et al., 2019) contains a subset of the corpus QuAC (Choi et al., 2018), another dataset for CQA. As in CoQA, each QuAC dialogue was produced by two crowd workers (one student and one teacher) and answers are spans extracted from a given piece of text. However, on the contrary to CoQA, the student does not see the text from which answers are taken. As CoQA, it contains unanswerable questions. CANARD was created by manually annotating a subset of QuAC: each question in CANARD was associated to one single out-of-context rewriting. The train/dev/test sets of CANARD respectively contain 5,571/3,418/31,538 questions. CANARD was used for evaluating the impact of QR on Question Answering models in Vakulenko et al. (2021).

3. CoQA with Question Rewriting (CoQAR)

CoQAR[1] was created from CoQA in a way that is analogous to how CANARD was created from QuAC. However, while CANARD was annotated using crowdsourcing, we decided to hire two specialized native-speakers annotators. Their task was to annotate original (*in-context*) questions from CoQA with at least two and at most three distinct *out-of-context* rewritings. To make sure that they understand what was expected, we ourself annotated a dialogue and provided it as an example. An example of conversation annotated by the annotators is provided in Table 1.

While annotators were told to preserve the meaning of the original sentences, they were also asked to paraphrase in their rewritings. As a results, these annotations contrast with those of CANARD, where the structure of the original question is usually preserved in the rewriting. In total, $4.1k$ conversations of CoQA train set were annotated as well as all 500 conversations of

	Number of rewritings				
	0	1	2	3	total
train	365	108	31,378	13,210	45,061
dev	9	0	37	7,937	7,983

Table 2: Number of questions depending on the number of rewritings.

the dev set. Since the test set of CoQA is not available, no conversation were annotated from it. The train and dev sets of CoQAR respectively contain 45k and 8k questions. Table 2 summarizes the number of questions that have 0,1,2 or 3 rewritings.

Overall, passages contains from 75 to 1079 words, with an average of 275. Conversation length distribution is displayed in Figure 1.

On average, out-of-context rewritings are longer (8.8 words) than the original questions (5.5 words); Figure 2 shows the question length distribution.

Most conversations were annotated by only one annotator, but 50 conversations were annotated by both. We relied on these conversations to analyse the annotations. We extracted two rewritings per question and per annotator and, using a pair of rewritings as references and the other as hypothesis, we computed the SacreBLEU score (Post, 2018) and the BERT-score (Zhang et al., 2020). SacreBLEU gives us an insight on the similarity of the surface form of rewritings, while BERT-score gives us an insight on the semantic similarity. We obtained a SacreBLEU score of 32.67 and a BERT-score of 90.22: this suggests that the rewritings have diverse surface form while being close in terms of meaning.

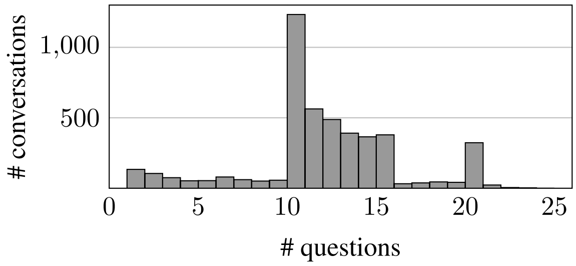

Figure 1: Distribution of conversations' length.

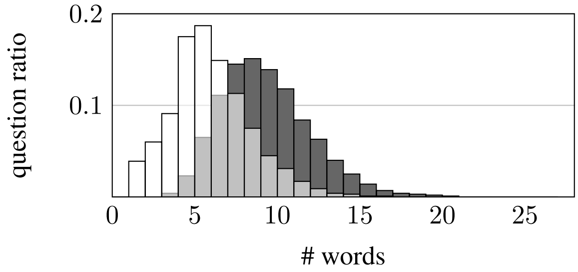

Figure 2: Distribution of length for original questions (white) and out-of-context rewritings (dark grey). Overlap of the distribution is light grey.

4. NLP Tasks

This section presents briefly the tasks of Question Paraphrasing (QP), Question Rewriting (QR) and Conversational Question Answering (CQA) that we used to evaluate the quality of the novel annotations of CoQAR.

4.1. Question Paraphrasing (QP)

QP is the task of transforming a source question into a question with equivalent meaning but different surface form (syntax, lexicon, etc.). In this paper we consider the case where both the source and paraphrased questions are out-of-context questions.

For each original question, CoQAR provides several out-of-context rewritings. We can regard two out-of-context rewritings of a same original in-context question as the source and paraphrase questions in the QP task.

We conducted experiments that consist in: (1) training QP models on CoQAR and an additional dataset, namely Quora Question Pairs (QQP); (2) evaluating the paraphrases generated by the models, via the standard metrics BLEU and METEOR, as well as human evaluation. More details about the experiments are presented in Section 5.1.

4.2. Question Rewriting (QR)

In QR, the model receives as input an in-context question, its conversational context, and the associated passage. Its task is to generate an out-of-context rewriting of the question.

We conducted the following experiment: (1) training QR models on CoQAR and CANARD; (2) evaluating these models, via standard metrics and human evaluation as presented in Section 5.2. Furthermore, we evaluate these QR models on downstream conversational question answering as presented in the next section and in Section 5.3.

4.3. Conversational Question Answering (CQA)

We consider CQA as a task for indirectly evaluating QR models. Typically, the inputs to a CQA neural model are: a question, its conversational context (i.e. the sequence of previous questions and answers), and the associated passage.

A challenge for conversational question answering was also released with CoQA[2]. The models are evaluated with the F1 score (Reddy et al., 2019). Transformers have been successfully used in this task: to the time this paper was written, the best model (a RoBERTa-based model (Ju et al., 2019)) got 90.7 of overall F1 measure, overcoming human performance 88.8.

[2]https://stanfordnlp.github.io/coqa/

Our goal is to indirectly assess the quality of QR by comparing the performance of a model taking original questions and their context as inputs with a model using out-of-context rewritings instead. In other words, we would like to know whether replacing the original question with its conversational context by the out-of-context rewriting has a positive impact on answer extraction. First, we evaluate the impact of rewritten questions in the performance of a RoBERTa baseline (Liu et al., 2019). Second, in order to assess the reusability of QR models trained on CoQAR, we further evaluate a state-of-the-art non-conversational QA model trained on SQuAD (Rajpurkar et al., 2018) by testing it with the rewritten questions. Please refer to Section 5.3 for more details about the evaluation of QR for this task.

5. Evaluation

In this section we present the settings and results of our experiments. Those involve the fine-tuning of the T5 and BART pretrained transformer models, with various training sets. We refer to fine tuned models with names of the form: "model(training-data-source)". For example, T5(CoQAR) will refer to a T5 model that was fine-tuned on data from CoQAR.

5.1. Question Paraphrasing

We first train QP models on CoQAR and Quora Question Pairs (QQP), then we evaluate the quality of the paraphrases generated by the models in terms of BLEU, METEOR and human evaluation.

Datasets: Each QP model was trained on a set of pairs consisting of a source question and its paraphrase, which are both out-of-context questions. We extracted such pairs from CoQAR and QQP. Since original questions of CoQAR have several out-of-context rewritings, we built pairs by associating rewritings of a same original question. This corresponds to a total of 237K paraphrase pairs, for an average of 1.9 paraphrase per out-of-context question. The QQP corpus is not a QA corpus: it was originally proposed as a Kaggle challenge to detect duplicate questions from Quora, a collaborative QA website where users can post their own questions or reply to those asked by others. The QQP corpus is composed of 404K question pairs, out of which 37 % are flagged as duplicates. We regard duplicate questions as paraphrases; assuming the transitivity of the semantic equivalence relation, clusters of paraphrases can be built. This results in a total of 710K paraphrase pairs, where each question is linked to 4.8 paraphrases on average. Clusters are partitioned into a training and test set with ratios 80 and 20 %, respectively.

Models: Three QP models are built by fine-tuning a pretrained BART model (Lewis et al., 2020) (*base* version[3]) on paraphrased question pairs. Each model is trained on one of three set of pairs: (1) pairs coming from CoQAR, (2) pairs coming from QQP, (3) pairs coming from both QQP and CoQAR. The models are fine-tuned during 2 epochs with batches of 10 samples. Optimization is done using AdamW, and static learning rate 5×10^{-5}.

Remark: Experiments with T5 models were also carried out but leading to slightly worse results. Thus, they are not reported here.

Test set	Model	BLEU	METEOR
CoQAR	Naive (copy)	0.694	0.492
	BART(CoQAR)	0.705	**0.537**
	BART(QQP)	0.673	0.464
	BART(CoQAR+QQP)	**0.737**	0.526
QQP	Naive (copy)	**0.737**	**0.634**
	BART(CoQAR)	0.626	0.445
	BART(QQP)	0.695	0.619
	BART(CoQAR+QQP)	0.692	0.611

Table 3: BLEU and METEOR scores on the test of CoQAR and QQP for various models. The Naive model simply outputs its input without any modification.

Objective Evaluation: Table 3 compares the BLEU and METEOR scores obtained by the fine-tuned BART models against a naive model that copies the input sentence as output. BLEU is provided for comparison purposes, even though it is known as less relevant for this task. Scores are measured on the test set of CoQAR and QQP.

First, the results show high values for the naive approach. This indicates (not surprisingly) that the source questions and their paraphrases are lexically close, especially in QQP. On CoQAR's test set, BART models whose training incorporates CoQAR data perform better than the naive model, demonstrating that fine-tuning enabled models to learn the task; on the other hand, on QQP's test set, the naive model gives the best results. These observations suggest that QQP may not be relevant for training and evaluating paraphrase generation models. Finally, we observe that using crossed data (training on CoQAR and testing on QQP, and vice versa) results quite logically in a loss of performance.

Human Evaluation: Two Mean Opinion Score (MOS) evaluations were carried out on 12 human testers who were asked to judge the quality of paraphrases. The objective is to complete observations from the automatic evaluations, as well as to study how CoQAR can benefit to the task on other datasets. We considered three corpora: CoQAR, QQP and CANARD. For each corpus, 50 source questions were randomly selected, and were paired with several paraphrases:

- one paraphrase from the corpus, to which we refer as the *reference*;
- one or several paraphrases generated by different

[3] https://huggingface.co/facebook/bart-base

Test set	Model	Meaning preservation		Linguistic correctness	
		MOS	(Std dev.)	MOS	(Std dev.)
CoQAR	Reference	3.82	(1.04)	4.46	(0.83)
	BART(CoQAR)	**3.97**	(1.15)	**4.54**	(0.75)
QQP	Reference	3.32	(1.28)	4.33	(1.01)
	BART(QQP)	**3.64**	(1.12)	4.37	(0.91)
	BART(CoQAR+QQP)	**3.65**	(1.21)	**4.51**	(0.66)
CANARD	BART(CoQAR)	**4.15**	(1.04)	**4.41**	(0.94)

Table 4: Results of the human evaluation of QP.

BART models: each source question from CoQAR and CANARD is paired with a paraphrase generated by BART(CoQAR), while each source question from QQP is paired with one paraphrase generated by BART(QQP) and one generated by BART(CoQAR+QQP).

In a first evaluation phase, testers were asked to judge the semantic similarity between two questions presented to them. Their opinion could be given on a 5-point scale: (1) "totally different"; (2) "mostly different"; (3) "half similar/half different"; (4) "mostly similar"; (5) "perfectly similar". In the second evaluation, each tester rated the linguistic correctness of single questions, independently of their meaning, on a similar scale to that used for semantic similarity. Each question pair (meaning preservation experiment) and single sentence (linguistic correctness) received 2 ratings.
Table 4 reports average values and standard deviation obtained for each MOS test. The main conclusions are given below, along with p-values from Mann-Withney U tests when relevant to assess the statistical significance between to mean values[4].
On CoQAR, paraphrases generated by BART obtain higher mean scores than the references, although the observed difference might be due to chance, both for meaning preservation ($p = 0.069$) and linguistic correctness ($p = 0.4$). This confirms that fine-tuning has indeed enabled the model to learn the task, as suggested by the BLEU and METEOR scores. On QQP also, BART models generalize well as they exceed references in terms of meaning preservation, although the difference might again be due to chance ($p = 0.091$). Adding CoQAR to the train set does not improve meaning preservation, and the slight increase in linguistic correctness is not statistically significant ($p = 0.37$). When comparing the second and last line of the table, it seems that the BART model learned on CoQAR transfers well to CANARD. However, it is possible that rewritings from CANARD are easier to paraphrase than those from CoQAR. Finally, is worth noting that QQP reference paraphrases obtained lower average meaning preservation scores than CoQAR paraphrases ($p = 0.019$). A manual investigation in QQP indeed shows that some questions are linked to more (or less) generic ones: for instance, "Given that C, what is A?" redirected to "What is A?", or "What is A?" redirected to "What are A and B?"). While this makes sense for helping users finding answers, these questions are not semantically equivalent. These observations suggest that QQP may not be relevant for training and evaluating paraphrase generation models.
Overall, the experiments demonstrate that CoQAR is conclusive to perform paraphrase generation on questions.

5.2. Question Rewriting

Test set	Model / train set	BLEU	METEOR
CoQAR	T5(CoQAR)	0.38	0.58
	T5(CANARD)	0.32	0.53
	T5(CoQAR+CANARD)	**0.39**	**0.59**
CANARD	T5(CoQAR)	0.31	0.57
	T5(CANARD)	**0.47**	**0.69**
	T5(CoQAR+CANARD)	0.44	0.66

Table 5: BLEU and METEOR scores obtained by the Question Rewriting models.

Datasets. For training and evaluation, we rely on CANARD and CoQAR. For CANARD, we use the original train/dev/test splits. For CoQAR, we use the original dev set as test set, and split the original train set into a train set and dev set, in such manner that CANARD and CoQAR dev sets have the same size. For training, we also make use of a mixture of CANARD and CoQAR, that we refer to as CoQAR+CANARD, whose train and dev sets are, respectively, the union of both corpora's train and dev sets. We train three variants of the QR model: one variant is trained on CANARD, one is trained on CoQAR, and the third one is trained on a mixture of both datasets.

Model: We train a QR model based on T5 on three datasets: CoQAR, CANARD, and CoQAR+CANARD. For each dataset, we fine-tune the small 1.1 version of T5[5]. We use AdamW optimizer, with initial learning rate 5×10^{-5} and no weight decay. After each epoch, the model is evaluated on the dev set

[4] As a reminder, the p-value measures the probability that the difference between two values is due to chance.

[5] `https://huggingface.co/google/t5-v1_1-small`

Test set	Model	Meaning preservation		Linguistic correctness	
		MOS	(Std dev.)	MOS	(Std dev.)
CoQAR	Human rewriting	**4.5**	(0.86)	**4.86**	(0.45)
	T5(CoQAR)	3.82	(1.42)	4.66	(0.82)
CANARD	Human rewriting	**4.60**	(0.96)	4.7	(0.89)
	T5(CANARD)	3.92	(1.34)	4.43	(1.08)
	T5(CoQAR+CANARD)	3.96	(1.47)	**4.76**	(0.77)

Table 6: Results of the human evaluation of QR.

using METEOR. We stop training as soon as the last obtained METEOR score is smaller than the two previous ones; we then keep the model that yielded the highest score.
Remark: BART models were also trained on the QR task; their BLEU and METEOR scores were overall similar but slightly worse than those of T5 models, thus we excluded them from the human evaluation phase and omitted them from the reported results.

Objective Evaluation Table 5 compares BLEU and METEOR scores obtained by the three fine-tuned T5 models. Scores are measured on CoQAR and CANARD test sets. Not surprisingly, performance drops when the models are tested on a data source which differs from the training data source (2st and 4th rows). On the contrary, mixing both corpora during training results in a unique model that performs well on both test sets (3rd and 6th rows). Scores are higher when testing on CANARD: this is again not surprising, since CoQAR rewritten questions have more diverse surface forms than those in CANARD, which are more similar to the original questions.

Human Evaluation Two Mean Opinion Score (MOS) evaluations were carried out on 8 human testers who were asked to judge the quality of rewritten questions. We sampled 50 original questions from CoQAR and 50 original questions from CANARD. Each original question was then paired with several rewritings:

- one rewriting from the corpus, to which we refer as the *reference*;
- one or several rewritings generated by different T5 models: each source question from CoQAR is paired with a rewriting generated by T5(CoQAR), while each source question from CANARD is paired with one rewriting generated by T5(CANARD) and one rewriting generated by T5(CoQAR+CANARD).

The pairs were then used in two evaluations.
In the first evaluation, rewritten questions were presented to human testers, together with the original question and its context (preceding dialogue turns and the corresponding text passage). Testers assessed the semantic similarity of the rewritten and original questions. In the second evaluation, rewritten questions were presented alone to the testers for them to assess linguistic correctness. Both semantic similarity and linguistic correctness were evaluated on the 5-points scale introduced in 5.1. In the end, each rewritten question received one rating for semantic similarity and one for linguistic correctness. The results are reported in Table 6.

QR mechanism	F1	EM
None (question+context)	**68.13**	**49.63**
Human rewriting	63.26	45.10
T5(CoQAR+CANARD)	63.30	44.97

Table 7: Results of the CQA evaluation.

We see that QR models obtain scores that are clearly below human performance in terms of meaning preservation. We also observe that the T5 model that was trained on CoQAR and CANARD obtains higher linguistic correctness scores than the model that was only trained on CANARD, and this result does not seem due to chance (a Mann-Whitney U test gives a p-value of 0.026). It is plausible that, although adding data from CoQAR to the training set does not improve meaning preservation, it improves linguistic correctness because of its greater diversity in term of rewritings' surface forms. Finally, note that the scores in Table 6 should not be compared with those of Table 4, because the sets of testers only partially overlap.

5.3. Conversational Question Answering

We would like to assess the impact of QR on state-of-the art models for CQA by answering the following question: would the models be able to extract the correct answer from the passage without dealing with the conversational context? To this aim we propose three experiments in which we train and evaluate a transformer on several variations of QR: no rewriting, human rewriting and model rewriting.

Datasets. We use CoQAR, with distinct rewriting.

i No rewriting: the orginal dataset, taking into account the conversational context.

ii Human rewriting: the dataset containing only the question rewritten by human annotators, ignoring completely the conversational context.

iii QR model: instead of using human annotations we use questions that were generated automatically by the T5(CoQAR+CANARD) model presented in Section 5.2.

Model. For the CQA experiments, we train and evaluate a RoBERTa[6] transformer on CoQAR with the distinct rewriting mechanisms described above. We fine tune the model during up to 5 epochs. We used Adam optimiser (Kingma and Ba, 2015), with learning rate of $5e-5$ and 12 gradient accumulation steps.

CQA evaluation. Results are presented in Table 7. Surprisingly, resolving the context with human question rewriting does not seem to help RoBERTa to better identify the answer in terms of F1 and exact match (EM) as defined in (Rajpurkar et al., 2016). We obtained an F1 and EM gain of 4.87 and 4.53 respectively of the original in-context questions over the out-of-context human rewritings.
Unlike Vakulenko et al. (2021), where results of the same task are reported on CANARD, the setting relying on original questions (referred to as CANARD_O) and the one relying on human-written questions (CANARD_H) respectively obtain 53.65 and 57.12 F1 scores, which correspond to a gain of 3.47 points for human rewriting. We suspect that the self-attention mechanism of RoBERTa solves the coreferences and ellipsis present in short in-context questions limited by the separation token from the context and the passage. While processing a long self-contained rewriting might be more difficult. These results confirm the good performance of RoBERTa on the original task of CQA (Ju et al., 2019).
Interestingly, automatically rewritten questions trained on both CoQAR and CANARD obtained similar performance than human rewritings, although human rewriting, got a slightly better EM. These results are comparable with the ones reported on CANARD in Vakulenko et al. (2021).

Reusability evaluation. To assess the reusability of the QR models trained on CoQAR, we compare the performances, on CoQAR and CANARD, of an existing QA model, with several question rewriting techniques, including QR. The considered QA model is the hugging-face *distilbert-base-uncased-distilled-squad*[7], which was trained on SQuAD. We adopted the same preprocessing as before: (i) no rewriting, (ii) human rewriting, (iii) QR model.
Table 8 shows that DistilBERT obtains higher F1 scores on CoQAR. For both test set, the best F1 scores are obtained when using human-rewritten questions. In terms of exact match, better results are obtained on CANARD: however, almost all exact matches are obtained on questions whose answer is "unknown". This could be explained by the fact that questions with unknown answers constitute about 18% of questions in CANARD, but less than 2% in CoQAR. Overall, it seems that the chosen QA model cannot handle CANARD correctly, independently on the QR step. On CoQAR, using human rewritings yields a significant increase of F1-score: from 35.90 F1 to 42.21. Interestingly, QR models produce results that come very close to human rewriting. Thus, the results on CoQAR suggest that the QR models are able, as a pre-processing step, to improve the results of simple QA systems on CQA.

6. Conclusion

In this paper, we presented CoQAR, a subset of CoQA where questions were annotated with out-of-context paraphrases. We took ethical concerns seriously, thus we hired two specialised native annotators for the task. Each question was annotated with several parapharses, and we demonstrated the richness of these paraphrases in terms of diversity in the surface form. Moreover, we evaluated the quality of the annotations via three tasks: QP, QR and CQA.
The results of the QP experiments suggest that CoQAR is more adapted to the task of Question Paraphrasing than QQP. Moreover, the human evaluation in Subsection 5.2 shows that the out-of-context rewritings of CoQAR are approximately as good as those of CANARD in terms of linguistic correctness and semantic similarity. This conclusion is also supported by the results of our experiments on QP and QR, where adding data from CoQAR to QQP or to CANARD during training does improve linguistic correctness.
Finally, although the results of our experiments confirm that QR performed either by humans or by models, does not improve the performance of CQA; it does enable the usage of non-conversational QA in CQA settings.

7. Acknowledgments

The work of recruiting and coordinating annotators was done by ELDA[8]. In particular, we would like to thank Khalid Choukri, Lucille Blanchard, and Marwa Hadj Salah. This work was fully funded by Orange Innovation. We thanks all the member of the DATA-AI department for their support. This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011011407 made by GENCI

8. Bibliographical References

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October. Association for Computational Linguistics.

[6] `https://huggingface.co/`
[7] `https://huggingface.co/distilbert-base-uncased-distilled-squad`
[8] `http://www.elra.info/en/about/elda/`

Test set	QR mechanism	F1	EM	unknown
CoQAR	None (question+context)	35.89	8.22	0.5
	Human rewriting	**42.21**	**9.23**	0.4
	T5(CoQAR)	41.56	9.09	0.4
	T5(CANARD)	39.84	8.84	0.4
	T5(CoQAR+CANARD)	41.80	9.17	0.4
CANARD	None (question+context)	27.46	**17.63**	16.9
	Human rewriting	**28.02**	16.08	15.3
	T5(CANARD)	27.21	15.76	15.7
	T5(CoQAR)	27.06	16.44	14.9
	T5(CoQAR+CANARD)	27.18	16.03	15.3

Table 8: Results of the reusability evaluation. The "unknown" column contains the percentage of questions with no answer where an exact match is obtained.

Elgohary, A., Peskov, D., and Boyd-Graber, J. (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 (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China, November. Association for Computational Linguistics.

Ju, Y., Zhao, F., Chen, S., Zheng, B., Yang, X., and Liu, Y. (2019). Technical report on conversational question answering. *arXiv preprint arXiv:1909.10772*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July. Association for Computational Linguistics.

Reddy, S., Chen, D., and Manning, C. (2019). CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, May.

Saha, A., Pahuja, V., Khapra, M., Sankaranarayanan, K., and Chandar, S. (2018). Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Usbeck, R., Ngomo, A.-C. N., Conrads, F., Röder, M., and Napolitano, G. (2018). 8th challenge on question answering over linked data (QALD-8). In *Joint proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018)*, volume 2241 of *CEUR Workshop Proceedings*, pages 51–57.

Vakulenko, S., Longpre, S., Tu, Z., and Anantha, R. (2021). Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 355–363.

Zhang, T., Kishore, V., Wu, F., Weinberger, K., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.