Table 1: Break-down analysis of the passage retrieval on TREC CAsT. Each row represents a group of QA samples that exhibit similar behaviour. We consider three types of input for every QA sample: the question from the test set (Original), generated by the best QR model (Transformer++) or rewritten manually (Human). The numbers correspond to the count of QA samples for each of the groups. The numbers in parenthesis indicate how many questions in the ground truth do not need rewriting, i.e., Human = Original.

			P@1	NDCG@3		
Original	QR	Human	= 1	> 0	≥ 0.5	= 1
×	×	×	49 (14)	10 (1)	55 (20)	154 (49)
✓	×	×	0	0	0	0
×	✓	×	2	0	1	0
✓	✓	×	0	1	1	0
×	×	✓	19	10	25	4
✓	×	✓	0	1	0	0
×	✓	✓	48	63	47	11
✓	✓	✓	55 (37)	88 (52)	44 (33)	4 (4)
			Total			173 (53)

Table 2: Break-down analysis of all reading comprehension results for the CANARD dataset, similar to Table 1. Observe that Table 1 is much more sparse than Table 2. There are almost no cases for which original or model-rewritten questions outperformed human rewriting when ranking passages. On the contrary, Table 2 indicates a considerable number of anomalous cases in which the reading comprehension model was able to answer incomplete follow-up questions but failed in answering ground-truth questions (rows 2-4).

Original	QR	Human	F1 > 0	F1 ≥ 0.5	F1 = 1
×	×	×	847 (136)	1855 (235)	2701 (332)
✓	×	×	174	193	181
×	✓	×	19	35 (2)	40 (1)
✓	✓	×	135	153	120
×	×	✓	141	288	232
✓	×	✓	65 (1)	57 (1)	40
×	✓	✓	226	324	269
✓	✓	✓	3964 (529)	2666 (428)	1988 (333)
			Total		5571 (666)

sion, which measures word overlap between the predicted answer span and the ground truth.

In contrast with QA, there is no established methodology for reporting the QR performance yet. Existing studies tend to report performance using BLEU metrics following the original CANARD paper (Yu et al., 2020; Lin et al., 2020).

We conducted a systematic evaluation of different performance metrics to find a subset that correlates with the human judgement of the quality of question rewrites (see more details in (Anantha et al., 2020)). Our analysis showed that ROUGE-1 Recall (ROUGE-1 R) (Lin, 2004) and Universal Sentence Encoder (*USE*) (Cer et al., 2018) embeddings correlate with the human judgement of the rewriting quality the most (Pearson 0.69 for ROUGE-1 R and Pearson 0.71 for USE). Therefore, we also use ROUGE-1 R and USE in this study to measure similarity between the model rewrites and the ground truth.

ROUGE is traditionally employed for evaluation of the text summarisation approaches. While ROUGE is limited to measuring lexical overlap between the two input texts, the USE model outputs dense vector representations that are designed to indicate semantic similarities between sentences beyond word overlap.

3 Our Error Analysis Framework

To trace how the difference in question formulation affects QA performance we compare the QR performance metrics with the QA performance metrics on the case-by-case basis, i.e., for each question-answer pair. Notice that we can apply the same approach for both retrieval and reading comprehension evaluation (see Table 2 for the results on reading comprehension).

Table 1 illustrates our approach. Each row of the table represents one of the combinations of the possible QA results for 3 types of question formulation. For every answer in a dataset we have 3 types of question formulation: (1) an original, possibly implicit, question (**Original**), (2) rewrites produced by the QR model (**QR**) and (3) rewrites produced by a human annotator (**Human**).

✓ indicates that the answer produced by the QA model was correct or × – incorrect, according to the thresholds provided in the right columns. For example, the first row of the table indicates the situation, when neither the original question, nor they generated or human rewrite were able to solicit the correct answer (× × ×). We then can automatically calculate how many samples in our results fall into each of these bins.

We assume human question rewrites as the ground truth. Therefore, all cases in which these rewrites did not result in a correct answer are errors of the QA component (rows 1-4 in Tables 1-2). The next two rows 5-6 show the cases, where human rewrites succeeded but the model rewrites failed, which we consider to be a likely error of the QR component. The last two rows are true positives for our model, where the last row combines cases where the original question was just copied without rewriting (numbers in brackets) and other cases

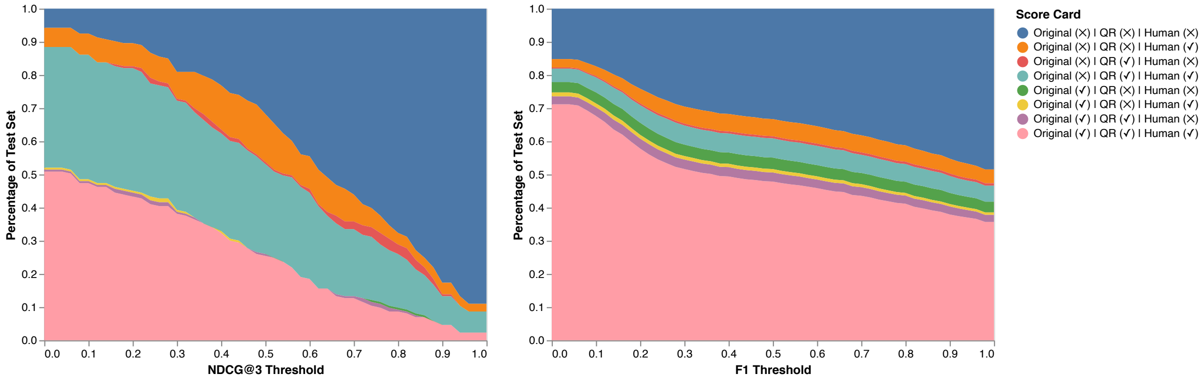

Figure 1: Break-down analysis for passage retrieval (left) and reading comprehension (right) results (best in color). This plot visualises the difference between the error distributions reported in Tables 1-2 with a sliding cut-off threshold. The blue region at the top of the plots represents the proportion of errors in QA. The orange region represents the proportion of errors in QR. The light green region shows the proportion of samples that were both rewritten and answered correctly. The pink region at the bottom shows the importance of QR for the task. The larger the region the more questions were answered correctly without any rewriting. We observe that the difference between the two plots is very pronounced. The impact of QR is noticeable in passage retrieval, which is more sensitive to question formulation than the reading comprehension model.

when rewriting was not required.

Since there is no single binary measure for the answer correctness, we can pick different cut-off thresholds for the QA metrics. For example, P@1=1 will consider the answer correct if it came up at the top of the ranking; or F1=1 will consider the answer correct only in cases with full span overlap, i.e, exact matches only. Figure 1 extends this analysis by considering all thresholds in the range [0; 1] with 0.02 intervals for NDCG@3 in retrieval and F1 in reading comprehension. This figure shows the proportion of different error types as well as the results sensitivity to the choice of the performance threshold.

4 Evaluation Results

Our approach allows us to estimate that the majority of errors stem from the QA model: 29% of the test samples for retrieval and 55% for reading comprehension. 11% of errors can be directly attributed to QR in the retrieval setup and 5% in reading comprehension.

To estimate the impact of QR on QA, we consider only the last four rows in Table 2 for which QA model return a correct answer for Human questions. Then, we divide the number of questions for which the Original question leads to the correct answer, i.e., without rewriting, by the total number of questions that can be answered correctly by our QA model (discarding the number of questions that were not rewritten by the annotators indicated in parenthesis). For example, to estimate the number of questions correctly answered in CANARD (F1 = 1) without rewriting, i.e., using the Original question as input to the QA model:

$$\frac{40 + 1988 - 333}{232 + 40 + 269 + 1988 - 333} = 0.77 \qquad (1)$$

The proportion of QR errors for retrieval setup is higher than for reading comprehension setup. In particular, we found that the majority of questions in CANARD test set (77% F1 = 1) can be correctly answered using only the Original questions without any question rewriting, i.e., even when the questions are ambiguous. For TREC CAsT, the chances of reaching the correct answer set using an ambiguous question are much lower (21% P@1 = 1). See Tables 3-4 for the complete result set with different cut-off thresholds.

There are two anecdotal cases where our QR component was able to generate rewrites that helped to produce better ranking than the human-written questions. The first example shows that the re-ranking model does not handle paraphrases well. Original question: "*What are good sources in food?*", human rewrite: "*What are good sources of **melatonin** in food?*", model rewrite: "*What are good sources in food for **melatonin***". In the second example the human annotator and our model chose different context to disambiguate the original question. Original question: "*What about environmental factors?*", human rewrite: "*What about environ-*

Table 3: The fraction of questions in TREC CAsT that were answered correctly without rewriting.

	P@1	NDCG@3		
Questions	$= 1$	> 0	≥ 0.5	$= 1$
All	0.45	0.55	0.38	0.21
Human != Original	0.21	0.34	0.13	0

Table 4: The fraction of questions in CANARD that were answered correctly without rewriting.

Questions	F1 > 0	F1 ≥ 0.5	F1 $= 1$
All	0.92	0.82	0.80
Human != Original	0.91	0.79	0.77

mental factors during the ***Bronze Age collapse****?*", model rewrite: "*What about environmental factors that lead to led to a* ***breakdown of trade***". Even though both model rewrites are not grammatically correct they solicited correct top-answers, while the human rewrites failed, which indicate flaws in the QA model performance.

5 QA-QR Correlation

In this section, we check whether the QR metrics can predict the QA performance for the individual questions by measuring the correlation between the QR and QA metrics. This analysis shows how the change in question formulation affects the answer selection. In other words, we are interested whether similar questions also produce similar answers, and whether distinct questions result in distinct answers.

To discover the correlation between question and answer similarity, we discarded all samples, where the human rewrites do not lead to the correct answers (top 4 rows in Tables 1-2). The remaining subset contains only the samples in which the QA model was able to find the correct answer. We then compute ROUGE for the pairs of human and generated rewrites, and measure its correlation with P@1 to check if rewrites similar to the correct question will also produce correct answers, and vice versa.

There is a strong correlation for ROUGE = 1, i.e., when the generated rewrite is very close to the human one. However, when ROUGE < 1 the answer is less predictable. Even for rewrites that have a relatively small lexical overlap with the ground-truth (ROUGE $\leq$ 0.4) it is possible to retrieve a correct answer, and vice versa.

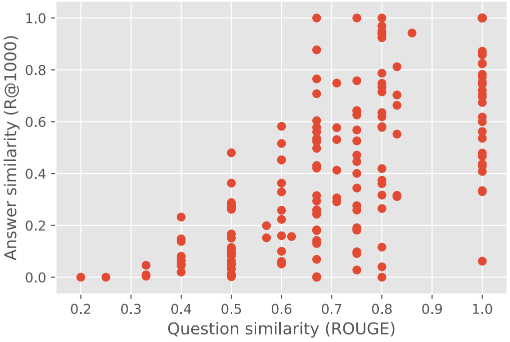

Figure 2: Strong correlation (Pearson 0.77) between question similarity (ROUGE) and the answer sets produced by the passage retrieval model (Recall).

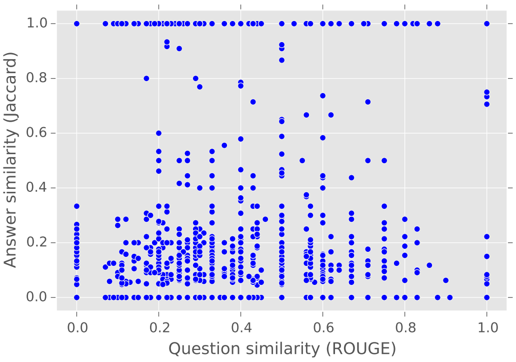

Figure 3: Weak correlation (Pearson 0.31) between question similarity (ROUGE) and answers produced by the reading comprehension model (Jaccard).

We further explore the effect of the QR quality on the QA results by comparing differences of the answer sets produced when given different rewrites. We compare answers produced separately for human and model rewrites using the same input question. However, this time we look at all the answers produced by the QA model irrespective of whether the answers were correct or not. This setup allows us to better observe how much the change in the question formulation triggers the change in the produced answer.

Figure 2 demonstrates strong correlation between the question similarity, as measured by ROUGE, and the answer set similarity. We measured the similarity between the top-1000 answers returned for the human rewrites and the generated rewrites by computing recall (R@1000). Points in the bottom right of this plot show sensitivity of the QA component, where similar questions lead to different answer rankings. The data points that are close to the top center area indicate weaknesses of