Question rewriting? Assessing its importance for conversational question answering *

Gonçalo Raposo[0000−0001−7806−6526] (✉), Rui Ribeiro[0000−0003−0922−1806], Bruno Martins[0000−0002−3856−2936], and Luísa Coheur[0000−0002−2456−5028]

INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
{goncalo.cascalho.raposo,rui.m.ribeiro,
bruno.g.martins,luisa.coheur}@tecnico.ulisboa.pt

Abstract. In conversational question answering, systems must correctly interpret the interconnected interactions and generate knowledgeable answers, which may require the retrieval of relevant information from a background repository. Recent approaches to this problem leverage neural language models, although different alternatives can be considered in terms of modules for (a) representing user questions in context, (b) retrieving the relevant background information, and (c) generating the answer. This work presents a conversational question answering system designed specifically for the Search-Oriented Conversational AI (SCAI) shared task, and reports on a detailed analysis of its question rewriting module. In particular, we considered different variations of the question rewriting module to evaluate the influence on the subsequent components, and performed a careful analysis of the results obtained with the best system configuration. Our system achieved the best performance in the shared task and our analysis emphasizes the importance of the conversation context representation for the overall system performance.

Keywords: Conversational Question Answering · Conversational Search · Question Rewriting · Transformer-Based Neural Language Models.

1 Introduction

Conversational question answering extends traditional Question Answering (QA) by involving a sequence of interconnected questions and answers [3]. Systems addressing this problem need to understand an entire conversation flow, often using explicit knowledge from an external datastore to generate a natural and correct answer for the given question. One way of approaching this problem is to divide it into 3 steps (see Fig. 1): initial question rewriting, retrieval of relevant information regarding the question, and final answer generation.

* Work supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), under project UIDB/50021/2020; by FEDER, Programa Operacional Regional de Lisboa, Agência Nacional de Inovação (ANI), and CMU Portugal, under project Ref. 045909 (MAIA) and research grant BI|2020/090; and by European Union funds (Multi3Generation COST Action CA18231).

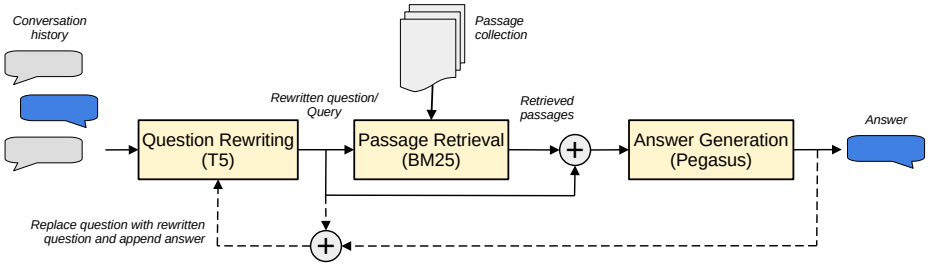


Fig. 1: Proposed conversational question answering system. Question rewriting is performed using T5, passage retrieval using BM25, and answer generation using Pegasus. Dashed lines represent different inputs explored for question rewriting.

In a conversational scenario, questions may contain acronyms, coreferences, ellipses, and other natural language elements that make it difficult for a system to understand the question. Question rewriting aims to solve this problem by reformulating the question and making it independent of the conversation context [5], which has been shown to improve systems performance [11].

After an initial understanding of the question and its conversational context, the next challenge is the retrieval of relevant information [14] to use explicitly in the answer generation [4]. For this step, the rewritten question is used as a query to an external datastore, and thus the performance of the initial rewriting module can affect the conversational passage retrieval [12].

The last module has the task of generating an answer that incorporates the retrieved information conditioned on the rewritten question. The Question Rewriting in Conversational Context (QReCC) dataset [1] brings these tasks together, supporting the training and evaluation of neural models for conversational QA. Although there are datasets for each individual task (e.g., CANARD for question rewriting [5] and TREC CAsT for passage retrieval [4]), to the best of our knowledge, QReCC is the only dataset that contemplates all these tasks.

This work presents a conversational QA system[1] implemented according to the dataset and task definition of the Search-Oriented Conversational AI (SCAI) QReCC 2021 shared task[2], specifically focusing on the question rewriting module. Participating as team *Rachael*, our system achieved the 1st place in this shared task. Besides evaluating the system performance as a whole, using many variations of the question rewriting module, our work highlights the importance of this module and how much it impacts the performance of subsequent ones.

2 Conversational Question Answering

To perform conversational question rewriting, the proposed system uses the model `castorini/t5-base-canard`[3] from the HuggingFace model hub [13]. This

[1] Available at `https://github.com/gonced8/rachael-scai`
[2] `https://scai.info/scai-qrecc/`
[3] `https://huggingface.co/castorini/t5-base-canard`

consists of a T5 model [8] which was fine-tuned for question rewriting using the CANARD dataset [5]. No further fine-tuning was performed with QReCC data.

In order to incorporate relevant knowledge when answering the questions, our system uses a passage retrieval module built with Pyserini [7], i.e., an easy-to-use Python toolkit that allows searching over a document collection using sparse and dense representations. In our implementation, the retrieval is performed using the BM25 ranking function [10], with its parameters set to $k_1 = 0.82$ and $b = 0.68$. This function is used to retrieve the top-10 most relevant passages.

Since our system needs to extract the most important information from the retrieved passages, which are often large, we used a Transformer model pre-trained for summarization. We chose the Pegasus model [15], more specifically, the version `google/pegasus-large`[4], which can handle inputs up to 1024 tokens.

We further fine-tuned the Pegasus model for 10 epochs in the task of answer generation, which can be seen as a summarization of the relevant text passages conditioned on the rewritten question. The training instances used the ground truth rewritten question concatenated with the ground truth passages (and additional ones retrieved with BM25), and the ground truth answers as the target.

3 Evaluation

3.1 Experimental Setup

The dataset used for both training and evaluation was the one used in the SCAI QReCC 2021 shared task, which is a slight adaption of the QReCC dataset. The training data contains 11 k conversations with 64 k question-answer (QA) pairs, while the test data contains 3 k conversations with 17 k questions-answer pairs. For each QA pair, we have also the corresponding truth rewrites and relevant passages, which are not considered during testing (unless specified otherwise).

To evaluate each module, we used the same automatic metrics as the shared task: ROUGE1-R [6] for question rewriting, Mean Reciprocal Rank (MRR) for passage retrieval, and F1 plus Exact Match (EM) [9] for the model answer evaluation. We additionally used ROUGE-L to assess the answer. When the system performs retrieval without first rewriting the question, we still report (between parentheses) the ROUGE1-R metric comparing the queries and truth rewrites.

3.2 Results

Question Rewriting Input We first studied different inputs to the question rewriting module in terms of the conversation history. Instead of using the original questions, one could replace them with the corresponding previous model rewrites. Moreover, one could use only the questions or also include the answers generated by the model. Regarding the length of the conversation history considered for question rewriting, we use all the most recent interactions that fit in the input size supported by the model.

[4] `https://huggingface.co/google/pegasus-large`

Table 1: Evaluation of multiple variations of the input used in the question rewriting module: Question (Q), Model Answer (MA), Model Rewritten (MR).

Description	Rewriting Input	Rewriting	Retrieval	Answer		
		ROUGE1-R	MRR	F1	EM	ROUGEL-F1
SCAI baseline: question	-	-	-	0.117	0.000	0.116
SCAI baseline: retrieved	-	(0.571)	0.065	0.067	0.001	0.073
SCAI baseline: GPT-3	-	-	-	0.149	0.001	0.152
No rewriting ($h = 1$)	-	(0.571)	0.061	0.136	0.005	0.143
No rewriting ($h = 7$)	-	(0.571)	0.145	0.155	0.003	0.160
Questions	(Q) + Q	0.673	**0.158**	0.179	**0.011**	0.181
Questions + answers	(Q + MA) + Q	0.681	0.150	0.179	0.010	0.181
Rewritten questions	(MR) + Q	0.676	0.157	0.187	0.010	0.188
Rewritten + answers	(MR + MA) + Q	**0.685**	0.149	**0.189**	0.010	**0.191**
Ground truth rewritten	-	(1)	0.385	0.302	0.028	0.293

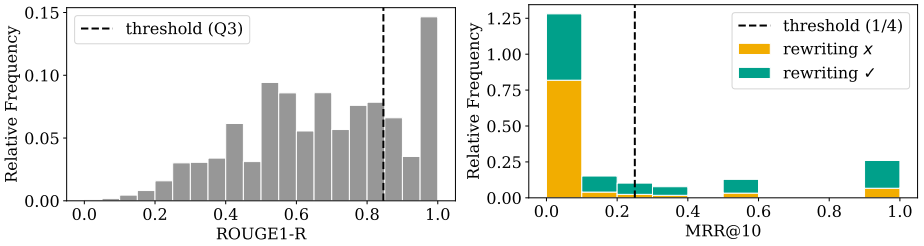
The results of our analysis are shown in Table 1, which also includes 3 baselines from the SCAI shared task[5]. The first baseline – question – uses the question as the answer; the second baseline – retrieval – uses the question to retrieve the top-100 most relevant passages using BM25, and selects the one with the highest score; the third baseline – GPT-3 – uses this Transformer Decoder [2] to generate the answer, prompting the model with an example conversation and the current conversation history. Among the baselines, GPT-3 achieved the best performance, which could be expected from this large language model. Moreover, the question baseline achieved better results than the retrieval baseline. This might be caused by the retrieved relevant passage being paragraph-like instead of conversational (thus, significantly different from the ground truth answer) since the performance doubled when we introduced the generation module.

Regarding our results, we observe that the variations without question rewriting had the worst performance, especially when only the last question is considered ($h = 1$). When introducing question rewriting, we explored 4 variations of the question rewriting input, all exhibiting higher scores than without question rewriting. In particular, the highest scores occur in 2 of the variations: when using only the questions, and when using both the model rewritten questions and model answers. The variation without model outputs in the question rewriting should be more resilient to diverging from the conversation topic.

When we used the ground truth rewritten questions instead, the performance of the passage retrieval and answer generation components increased about $1.6 \sim 2.5\times$, highlighting the importance of good question rewriting.

Impact of Question Rewriting After this initial evaluation, we used the system with the highest F1 score (rewriting using model rewritten questions and answers) to further evaluate the impact of question rewriting. We computed the aforementioned metrics for each QA pair and used the scores to classify the results into different splits reflecting result quality, allowing us to analyze a module's performance when the previous ones succeeded (✓) or failed (✗).

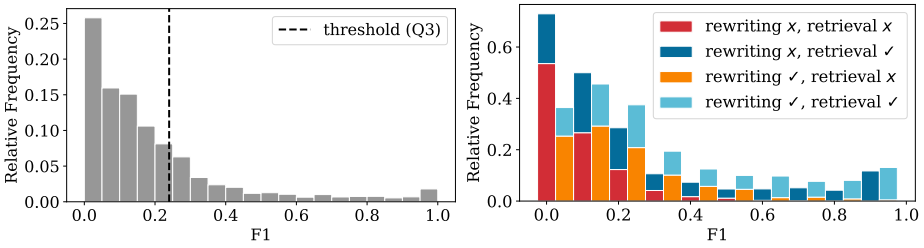
[5] `https://www.tira.io/task/scai-qrecc`



(a) Distribution of ROUGE1-R scores for question rewriting.

(b) Distribution of MRR scores (retrieval) when question rewriting succeeds or fails.

Fig. 2: Analysis of the influence of question rewriting on passage retrieval performance. Relative frequencies refer to the number of QA pairs of each split.



(a) Distribution of F1 scores for the answer generation component.

(b) Distribution of F1 scores when rewriting and retrieval succeed and fail.

Fig. 3: Analysis of the influence of question rewriting and passage retrieval on answer generation performance. Relative frequencies refer to each split.

To classify the performance of the question rewriting module using ROUGE scores, we used the 3rd quartile of the score distribution as a threshold (shown in Fig. 2a), since we are unable to choose a value that corresponds exactly to right/wrong rewriting decisions. As for classifying the passage retrieval using the MRR score, an immediate option would be to classify values greater than 0 as successful. However, although our system retrieves the top-10 most relevant passages, the answer generation model is limited by its maximum input size, which resulted in less important passages being truncated. A preliminary analysis showed us that, in most QA pairs, the model only considered $3 \sim 4$ passages, and therefore we defined the threshold of a successful retrieval as MRR $\geq 1/4$.

When the question rewriting succeeds (ROUGE1-R $\geq$ Q3), the passage retrieval also exhibits better performance, as seen by MRR scores greater than 0 being more than twice more frequent (see Fig 2b). Although both splits have many examples where the retrieval fails completely (MRR $= 0$), they are about twice more frequent when the question rewriting fails.

Fig. 3a shows the distribution of F1 scores for answer generation, revealing that 75 % of the results have an F1 score lower than 0.25. In turn, Fig. 3b shows 4 splits for when the question rewriting and retrieval modules each succeed or fail. Comparing the stacked bars together, one can analyze the influence of question rewriting in the obtained F1 score. Independently of the retrieval performance,

F1 scores higher than 0.2 are much more frequent when the rewriting succeeds than when it fails. In particular, F1 scores between 0.3 and 0.8 are about 2× more frequent when the rewriting succeeds. Moreover, poor rewriting performance results in about 2× more results with an F1 score close to 0. Analyzing in terms of MRR, higher F1 scores are much more frequent when the retrieval succeeded. Interestingly, if the rewriting fails but the retrieval succeeds (less probable, as seen in Fig. 2b), the system is still able to generate answers with a high F1 score.

Error Example In Table 2, we present a representative error where the system achieves a high ROUGE1-R score in the rewriting module but fails to retrieve the correct passage and to generate a correct answer. The only difference between the model and truth rewritten questions is in the omitted first name *Ryan*, which led the system to retrieve a passage referring to a different person (*Michael Dunn*). Although the first name was not mentioned in the context, maybe by enhancing the question with information from the previous turn (e.g., the age or day of death) the system could have performed better in the subsequent modules.

Table 2: Example conversation where the retrieval and generation failed.

Context		Q: When was Dunn's death? A: Dunn died on August 12, 1955, at the age of 59.
Question		What were the circumstances?
Rewriting	Truth	What were the circumstances of Ryan Dunn's death?
ROUGE1-R: 0.889	Model	What were the circumstances of Dunn's death?
Retrieval	Truth	`http://web.archive.org/web/20191130012451id_/https://en.wikipedia.org/wiki/Ryan_Dunn_p3`
MRR: 0	Model	`https://frederickleatherman.wordpress.com/2014/02/16/racism-is-an-insane-delusion-about-people-of-color/?replytocom=257035_p1`
Generation	Truth	Ryan Dunn's Porsche 911 GT3 veered off the road, struck a tree, and burst into flames in West Goshen Township, Chester County, Pennsylvania.
F1: 0.051, EM: 0, ROUGEL-F1: 0.128	Model	The Florida Department of Law Enforcement concluded that Dunn's death was a homicide caused by a single gunshot wound to the chest.

4 Conclusions and Future Work

This work presented a conversational QA system composed of 3 modules: question rewriting, passage retrieval, and answer generation. The results obtained from its evaluation on the QReCC dataset show the influence of each individual module in the overall system performance, and emphasize the importance of question rewriting. When the question rewriting succeeded, both the retrieval and answer generation improved – lower scores were up to 2× less frequent while higher scores were also about 2× more frequent. Future work should explore how to better control the question rewriting and its interaction with passage retrieval. Moreover, the impact of question rewriting or the use of other input representations should be validated with different datasets and models. Although our system with automatic question rewriting achieved the 1st place in the SCAI QReCC shared task, significant improvements can perhaps still be achieved with a better rewriting module (e.g., by fine-tuning T5 in the QReCC dataset).

References

1. Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., Chappidi, S.: Open-domain question answering goes conversational via question rewriting. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 520–534. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.44, https://aclanthology.org/2021.naacl-main.44
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
3. Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.t., Choi, Y., Liang, P., Zettlemoyer, L.: QuAC: Question answering in context. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2174–2184. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1241, https://aclanthology.org/D18-1241
4. Dalton, J., Xiong, C., Callan, J.: Cast 2020: The conversational assistance track overview. In: The Twenty-Ninth Text REtrieval Conference(TREC 2020) Proceedings (2020), https://trec.nist.gov/pubs/trec29/trec2020.html
5. Elgohary, A., Peskov, D., Boyd-Graber, J.: 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. pp. 5918–5924. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1605, https://aclanthology.org/D19-1605
6. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
7. Lin, J., Ma, X., Lin, S.C., Yang, J.H., Pradeep, R., Nogueira, R.: Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2356–2362. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463238
8. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
9. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1264, https://aclanthology.org/D16-1264

10. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009). https://doi.org/10.1561/1500000019
11. Vakulenko, S., Longpre, S., Tu, Z., Anantha, R.: Question rewriting for conversational question answering. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. p. 355–363. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3437963.3441748
12. Vakulenko, S., Voskarides, N., Tu, Z., Longpre, S.: A comparison of question rewriting methods for conversational passage retrieval (Jan 2021)
13. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (October 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, `https://www.aclweb.org/anthology/2020.emnlp-demos.6`
14. Yu, S., Liu, Z., Xiong, C., Feng, T., Liu, Z.: Few-shot conversational dense retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 829–838. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462856
15. Zhang, J., Zhao, Y., Saleh, M., Liu, P.: PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 11328–11339. PMLR (13–18 Jul 2020), `https://proceedings.mlr.press/v119/zhang20ae.html`