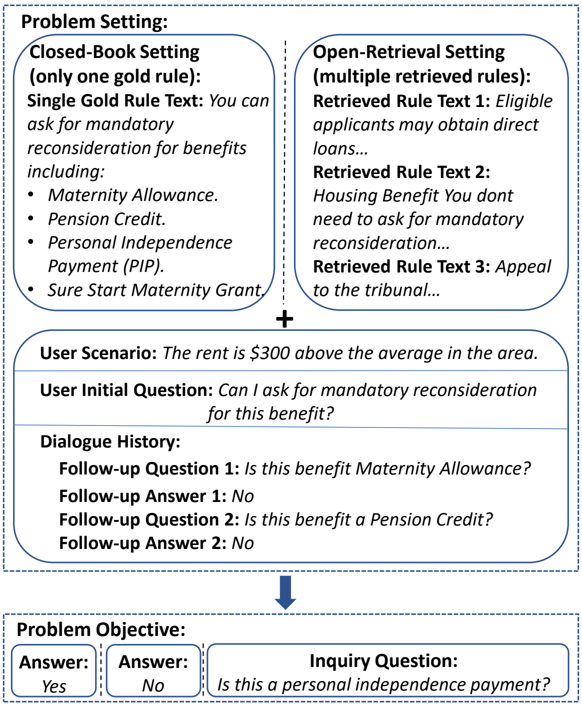
Towards End-to-End Open Conversational Machine Reading

Sizhe Zhou[1,3], Siru Ouyang[2,3], Zhuosheng Zhang[2,3] Hai Zhao[2,3, *]
[1] UM-SJTU Joint Institute, Shanghai Jiao Tong University
[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[3] Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University
{sizhezhou,oysr0926,zhangzs}@sjtu.edu.cn,zhaohai@cs.sjtu.edu.cn

Abstract

In open-retrieval conversational machine reading (OR-CMR) task, machines are required to do multi-turn question answering given dialogue history and a textual knowledge base. Existing works generally utilize two independent modules to approach this problem's two successive sub-tasks: first with a hard-label decision making and second with a question generation aided by various entailment reasoning methods. Such usual cascaded modeling is vulnerable to error propagation and prevents the two sub-tasks from being consistently optimized. In this work, we instead model OR-CMR as a unified text-to-text task in a fully end-to-end style. Experiments on the ShARC and OR-ShARC dataset show the effectiveness of our proposed end-to-end framework on both sub-tasks by a large margin, achieving new state-of-the-art results. Further ablation studies support that our framework can generalize to different backbone models.



Problem Setting:

Closed-Book Setting (only one gold rule):
Single Gold Rule Text: *You can ask for mandatory reconsideration for benefits including:*
- *Maternity Allowance.*
- *Pension Credit.*
- *Personal Independence Payment (PIP).*
- *Sure Start Maternity Grant.*

Open-Retrieval Setting (multiple retrieved rules):
Retrieved Rule Text 1: *Eligible applicants may obtain direct loans…*
Retrieved Rule Text 2: *Housing Benefit You dont need to ask for mandatory reconsideration…*
Retrieved Rule Text 3: *Appeal to the tribunal…*

+

User Scenario: *The rent is $300 above the average in the area.*

User Initial Question: *Can I ask for mandatory reconsideration for this benefit?*

Dialogue History:
Follow-up Question 1: *Is this benefit Maternity Allowance?*
Follow-up Answer 1: *No*
Follow-up Question 2: *Is this benefit a Pension Credit?*
Follow-up Answer 2: *No*

Problem Objective:

Answer: *Yes*
Answer: *No*
Inquiry Question: *Is this a personal independence payment?*

Figure 1: CMR and OR-CMR Task Overview

1 Introduction

In a multi-turn dialogue comprehension scenario, machines are expected to answer high-level questions through interactions with human beings until enough information is gathered to derive a satisfying answer (Zhu et al., 2018; Zhang et al., 2018; Zaib et al., 2020; Huang et al., 2020; Fan et al., 2020; Gu et al., 2021). As a specific and challenging dialogue comprehension task, conversational machine reading (CMR) (Saeidi et al., 2018) requires machines to understand the given user's initial setting and dialogue history before the machine itself is able to give a final answer or inquire for more clarifications according to rule texts (see Figure 1).

In terms of acquisition of rule texts which are the main reference for tackling the CMR, there is closed-book setting where the rule texts are all given and there is correspondingly open-retrieval setting where the rule texts need to be retrieved from a knowledge base (Gao et al., 2021) (see Figure 1). In terms of problem objectives, current approaches in general divide the targets into two categories, one as decision making and one as question generation (Zhong and Zettlemoyer, 2019; Lawrence et al., 2019; Verma et al., 2020; Gao et al., 2020b,c) . For decision making sub-task, the machine is required to give decisions to directly answer the user question which concludes the dialogue or generate clarifying questions which continues the dialogue. For question generation sub-task, the machine is required to generate the clarifying questions that are essential to the later final decision making. Following this line of approaching the CMR task, a variety of works

[*] Corresponding author. This paper was partially supported by Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

Findings of the Association for Computational Linguistics: EACL 2023, pages 2064–2076
May 2-6, 2023 ©2023 Association for Computational Linguistics

have been proposed mainly based on modeling the matching of elementary conditions (Henaff et al., 2017; Zhong and Zettlemoyer, 2019; Lawrence et al., 2019; Verma et al., 2020; Gao et al., 2020b,c; Ouyang et al., 2021; Zhang et al., 2021) in either a sequential encoding or graph-based manner.

However, by tackling the CMR task with two divided sub-tasks, the corresponding division of the optimization on decision making sub-task and the optimization on the question generation sub-task may result in problems including error propagation, thus hindering further performance advance. Ouyang et al. (2021) has shown that transferring some knowledge between the training of two sub-tasks is beneficial for better performance. However, reducing the gap between two sub-tasks to achieve an end-to-end optimization CMR task still needs further and more comprehensive attempts.

In this work, we propose a completely Unified end-to-end framework for Conversational Machine Reading tasks (UniCMR[1]) to tackle the division of optimization challenge by formulating the CMR/OR-CMR task into a single text-to-text task. Our contributions are summarized as follows:

(i) We completely unify two sub-tasks of OR-CMR into a single task in terms of optimization, achieving a fully end-to-end optimization paradigm.

(ii) Experimental results on the OR-ShARC dataset and ShARC dev set show the effectiveness of our method, especially on the question generation sub-task with a relatively small amount of parameters. Furthermore, our method achieves the new state-of-the-art results on all sub-tasks.

(iii) By further ablation studies, we have shown that our proposed framework largely advances the decision making performance, and reduces error propagation thus boosting the question generation performance. We have also shown that our proposed framework can generalize to different backbone models. Qualitative analysis including case study has further verified the effectiveness of our framework.

2 Related Work

2.1 Conversational Machine Reading

The mainstream of research on the conversation-based reading comprehension task focuses on either the decision making (Choi et al., 2018; Reddy et al., 2019; Sun et al., 2019; Tao et al., 2019; Cui et al., 2020; Yang et al., 2020) or the follow-up utterance generation (Wu et al., 2019; Bi et al., 2019; Ren et al., 2019; Gao et al., 2020a). However, the decision making centered approaches leave out cultivating the machine's capability to reduce the information gap by clarifying interactions. While the question generation centered approaches neglect exploring the machine's capability to concentrate on target-oriented information and make vital decisions. In contrast, our work focuses on a more challenging conversation-based reading comprehension task called conversational machine reading (CMR) task (Saeidi et al., 2018; Gao et al., 2021), which requires machines to make decisions and generate clarifying questions in a dialogue given rule texts and user scenarios.

2.2 Open-Retrieval CMR

Most of the current studies on CMR concentrate on the closed-book setting of CMR where the essential reference for the final decision, a piece of rule text corresponding to each dialogue, is given (Zhong and Zettlemoyer, 2019; Verma et al., 2020; Gao et al., 2020b,c). One typical example benchmark is called ShARC (Saeidi et al., 2018). However, in a more realistic and also more challenging setting, the machine is required to retrieve rule texts based on different scenarios. Similar to the open domain question answering setting where the supporting texts are retrieved from external documents to answer factoid questions (Moldovan et al., 2000; Voorhees and Tice, 2000), open-retrieval conversational machine reading (OR-CMR) task is established by requiring the machine to retrieve useful information from a given knowledge base composed of rule texts. In contrast to most of the previous works on CMR, we focus on OR-CMR in pursuit of a more realistic and more challenging setting.

2.3 Joint Optimization of CMR

Existing studies generally approach conversational machine reading task by separating it into two sub-tasks (Zhong and Zettlemoyer, 2019; Verma et al., 2020; Gao et al., 2020b,c), decision making and question generation. Therefore, existing approaches generally focus on different methods to extract the fulfillment of rule-related conditions and conduct explicit entailment reasoning on tracking the conditions in the dialogues. This includes applying attention mechanisms on the sequentially

[1] Our source codes are available at https://github.com/KevinSRR/UniCMR.

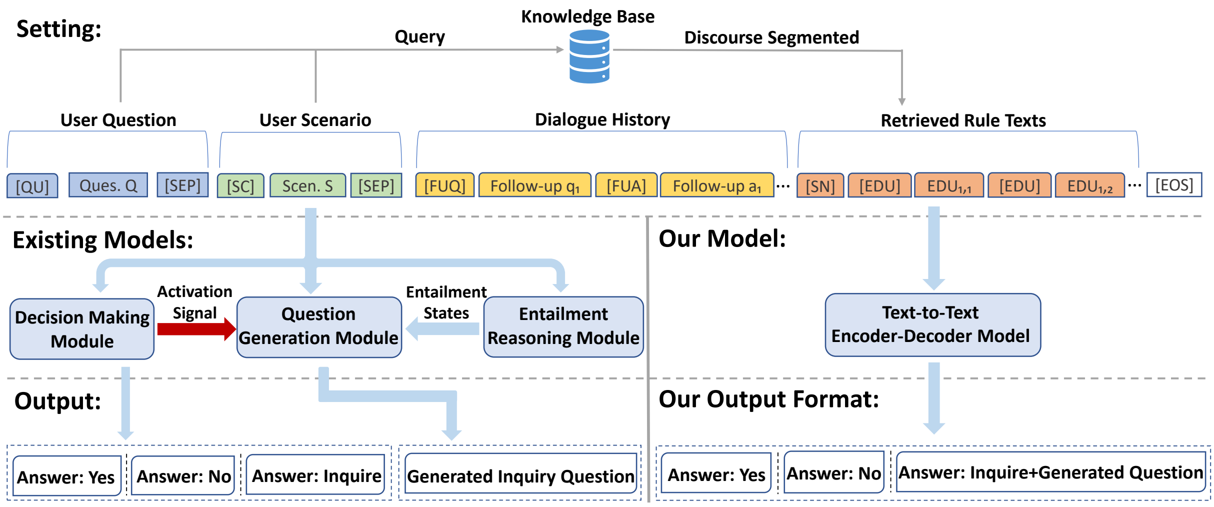


Figure 2: The overall framework for our proposed model (bottom right part) compared with the existing ones (bottom left part). Note that the ways of preprocessing the problem setting input vary from model to model, but they are generally similar. And the setting part only shows our preprocessing overview. Also note that [QU], [SEP], [SC], [SEP], [FUQ], [FUA], [SN], [EDU] are added special tokens while [EOS] is the end-of-sequence token for encoder-decoder model.

encoded user setups and the dialogue (Zhong and Zettlemoyer, 2019; Lawrence et al., 2019; Verma et al., 2020; Gao et al., 2020b,c) and extract discourse structures for better fulfillment matching (Ouyang et al., 2021).

However, one of the major challenges emerges as the division of the optimization of decision making sub-task and the optimization of the question generation sub-task. Zhang et al. (2021) have taken the initial attempt to mitigate the division between two sub-tasks by considering the encoded hidden states from decision making in question generation module. However, it still lacks synergy of optimization and relies on separate feature extractions including the entailment reasoning. In contrast, our work approaches the conversational machine reading task by unifying the two sub-tasks into one, enabling an end-to-end joint model optimization on both the decision making target and question generation target.

3 Problem Formulation

As shown by Figure 1, in traditional CMR task, the machine will be given: user scenario S, user initial question Q, a gold rule text R, and dialogue history $D := \{(q_1, a_1), (q_2, a_2), \dots, (q_n, a_n)\}$ which consists of n follow-up question-answer pairs. The machine is required to do the two sub-tasks:

- **Decision Making**. The machine makes a decision to either answer the user initial question with *Yes* or *No*, or give *Inquire* [2] which activates the second sub-task to generate the inquiry question for more clarification.
- **Question Generation**. The machine generates an inquiry question aimed at essential clarifications to answer the user's initial question.

Beyond CMR, open-retrieval conversational machine reading (OR-CMR) (Gao et al., 2021) further mimics the more challenging second scenario, which is the focus of this work. As shown by Figure 1, the difference between the CMR and OR-CMR lies in the rule text part R. In CMR, the machine is provided with a gold rule text in a closed-book style. While in OR-CMR, the machine needs to retrieve rule texts from a knowledge base in an open-retrieval style alternatively. The machine is given a knowledge base B containing rule texts. Therefore, under the OR-CMR setting, the machine needs to first retrieve m rule texts $R_1, R_2, \dots, R_m$ to complete the input for the same downstream decision making and question generation sub-tasks.

4 Framework

Our model is composed of two main modules: a retriever and a text-to-text encoder-decoder model.

[2] For the completeness of the conversational machine reading task, there is an additional decision making answer *Irrelevant* which states that the user question is unanswerable. This is the case for CMR task. However, in our work, we mainly follow the setting of OR-CMR and assume that no such answer will be encountered.

The retriever is applied to retrieve rule texts R_1, $R_2, \ldots, R_m$ from a given knowledge base B. The text-to-text encoder-decoder model will take in the preprocessed textual input and generate the textual answer directly as a whole. Subsequent extraction methods will be applied for decision making and question generation sub-tasks to obtain the predictions for each sub-task respectively.

4.1 Retriever

To obtain the rule texts, the user scenario S and user initial question Q are concatenated as the input query to the retriever. Our retriever employs the MUDERN TF-IDF-based method (Gao et al., 2021), which takes account of bigram features and scores the similarity between rule texts and queries in the form of bag-of-words vectors weighted by the TF-IDF model. Top-scored m rule texts R_1, R_2, …, R_m will then be chosen for the following text-to-text encoder-decoder model.

4.2 Text-to-Text Encoder-Decoder

One of the major challenges of the CMR or OR-CMR task is the division of sub-task optimizations. Motivated by T5 (Raffel et al., 2020) which formulate several traditional NLP tasks into a unified text-to-text generation task, we unify the two sub-tasks by formulating the input and output to our encoder-decoder model as follows.

4.2.1 Input Formulation

Discourse Segmentation. We employ the discourse segmentation approach (Shi and Huang, 2019) to parse the retrieved rule texts into explicit conditions for the model. After discourse segmentation, each retrieved R_i is parsed into N_i elementary discourse units (EDUs) $EDU_{i,1}, EDU_{i,2}, \ldots, EDU_{i,N_i}$. Formulation of the final input I is shown by the setting part in Figure 2.

4.2.2 Output Formulation

The output of the text-to-text encoder-decoder will be a sequence of textual tokens O :=$\{o_1$, o_2, …, $o_k\}$ where the length k is determined by the model itself but within the maximum generation length hyperparameter. To extract the prediction of the decision making sub-task and the question generation sub-task respectively, we assume the first output token o_1 is model's prediction, and the following tokens $\{o_2$, …, $o_k\}$ are the generated follow-up question, which is only meaningful when o_1 represents the *Inquire* decision.

4.2.3 Training Objective

In training stage, the labels $Y := \{y_1, y_2, \ldots, y_k\}$ are formulated as: {*Yes* Token, [EOS]}, { *No* Token, [EOS]}, and {*Inquire* Token, Follow-up Question Tokens, [EOS]}.[3] The training objective is defined as:

$$\mathcal{L} = -\sum_{j=1}^{k} \log P(y_j | y_{<j}, I; \theta), \qquad (1)$$

where I is the input to our encoder-decoder model and θ is all the parameters of our model.

5 Experiments

5.1 Experiment Setups

Datasets. Our training and evaluation is based on the OR-ShARC dataset (Gao et al., 2021). Original dataset ShARC (Saeidi et al., 2018) contains 948 dialogues trees which is then flattened into 32,436 examples with entries composed of rule documents, user setups, dialogue history, evidence, and decision. Derived from ShARC, OR-ShARC modifies the *initial question* to be self-contained and to be independent of gold rule texts. Then the gold rule texts are removed to form the knowledge base B of 651 rules. The train and dev set of ShARC are further split into train, dev, and test set, with sizes 17,936, 1,105, and 2,373, respectively.

The dev and test set each satisfies that around 50% of examples ask questions based on the rule texts used in training (seen) and the remaining asks questions based on the unseen rule texts in training. This feature of the datasets aims to mimic more realistic scenario where user may asks questions on information that the machine has encountered or has never encountered (Gao et al., 2021).

Evaluation Metrics. For decision making sub-task, the evaluation is Micro- and Macro- Accuracy of the decisions. For question generation sub-task, we adopt the F1_{BLEU} (Gao et al., 2021) which calculates the F1 score with precision of BLEU (Papineni et al., 2002) when the predicted decision is *Inquire* and recall of BLEU when the ground truth decision is *Inquire*.

[3] To make sure *Yes* Token, *No* Token and *Inquire* Token have the same length after tokenization, we set the valid tokens of "1", "2" and "3" to serve as *Yes* Token, *No* Token and *Inquire* Token respectively without loss of generality.

Model	Dev Set				Test Set			
	Decision Making		Question Generation		Decision Making		Question Generation	
	Micro	Macro	F1$_{\text{BLEU1}}$	F1$_{\text{BLEU4}}$	Micro	Macro	F1$_{\text{BLEU1}}$	F1$_{\text{BLEU4}}$
w/ DPR++								
MUDERN	79.7±1.2	80.1±1.0	50.2±0.7	42.6±0.5	75.6±0.4	75.8±0.3	48.6±1.3	40.7±1.1
OSCAR	**80.5**±**0.5**	**80.9**±**0.6**	51.3±0.8	43.1±0.8	76.5±0.5	76.4±0.4	49.1±1.1	41.9±1.8
w/ TF-IDF								
E^3	61.8±0.9	62.3±1.0	29.0±1.2	18.1±1.0	61.4±2.2	61.7±1.9	31.7±0.8	22.2±1.1
EMT	65.6±1.6	66.5±1.5	36.8±1.1	32.9±1.1	64.3±0.5	64.8±0.4	38.5±0.5	30.6±0.4
DISCERN	66.0±1.6	66.7±1.8	36.3±1.9	28.4±2.1	66.7±1.1	67.1±1.2	36.7±1.4	28.6±1.2
DP-RoBERTa	73.0±1.7	73.1±1.6	45.9±1.1	40.0±0.9	70.4±1.5	70.1±1.4	40.1±1.6	34.3±1.5
MUDERN	78.4±0.5	78.8±0.6	49.9±0.8	42.7±0.8	75.2±1.0	75.3±0.9	47.1±1.7	40.4±1.8
UNICMR$_{\text{base}}$	75.6±0.4	76.5±0.6	53.7±0.5	46.5±0.2	71.7±1.2	72.2±1.1	48.4±1.5	41.5±1.7
UNICMR$_{\text{large}}$	77.7±0.5	78.0±0.6	**59.3±1.2 (↑8.0)**	**52.8±0.9 (↑9.7)**	**76.7±1.2 (↑0.2)**	**76.7±1.1 (↑0.3)**	**54.2±1.4 (↑5.1)**	**47.9±1.6 (↑6.0)**

Table 1: Results on the validation and test set of OR-ShARC. Numerical values in the parentheses show how much our proposed model outperforms the current SOTA model. The first block presents the results of public models with the DPR++ retrieval method, and the second block reports the results of TF-IDF retrieval-based public models and our SOTA model. Our average results with a standard deviation on 3 random seeds are reported. The numbers in brackets (↑) indicate the improved accuracy over the previous state-of-the-art model.

Model	Dev Set			
	Decision Making		Question Gen.	
	Micro	Macro	BLEU1	BLEU4
OSCAR	70.1	75.6	63.3	48.1
UNICMR	**72.6**	**78.0**	**66.3**	**53.9**

Table 2: Results on the validation set of ShARC (with large models). Note that the test set of ShARC is not public hence only the evaluation on dev set is conducted.

Implementation Details. Following the MUDERN model, we employ T5 as our text-to-text encoder-decoder model and initialize the model with the pretrained T5-base and T5-large weights[4]. For the main model either base or large, we set the max generation length as 30, number of beams in generation as 5, and use the first 8 top scored retrieved rule texts in preparing input. The training process utilizes AdamW (Loshchilov and Hutter, 2017) optimizer for 16 epochs with a learning rate of 3e-5. Max gradient norm of 1 is used to conduct gradient clipping. The batch size is 4 with a gradient accumulation step as 8. Random seeds 19, 27, and 95 are applied. Experiments are conducted in two RTX TITAN GPU's with 24G memory [5]. In training stage, the model with best F1$_{\text{BLEU4}}$ score on dev set is kept.

5.2 Quantitative Results

The effectiveness of our proposed method is verified on both the OR-ShARC and the original ShARC datasets. In addition, we compare the number of parameters with related studies. Tables 1-3 present our main experimental results. We will discuss our findings in the following part.

5.3 Decision Making and Question Generation performance on OR-ShARC.

Referring to our results reported in Table 1, our large unified model has achieved new SOTA question generation performance in both dev and test sets by a large margin. In terms of decision making results, our large model lags behind in the dev set but prevails in the test set performance by maintaining a stable and consistent performance when transferring from dev set to test set.

5.4 Performance on ShARC.

As a reference, the performance of the UNICMR$_{\text{large}}$ together with the current SOTA model OSCAR on the dev set of ShARC is reported on Table 2. Note that, in contrast with OR-ShARC (Gao et al., 2021), ShARC benchmark (Saeidi et al., 2018) is in the closed-book setting with the evaluation metric of the question generation sub-task as BLEU. Based on the results in Table 2, it can be seen that UNICMR$_{\text{large}}$ maintains a new SOTA performance on dev set by a large margin for both the decision making and the question generation sub-tasks. This shows our unified method is effective for the model's performance

[4] https://huggingface.co/t5-base, and https://huggingface.co/t5-large, respectively.

[5] Average training run time for UNICMR$_{\text{large}}$ is approximately 12 hours with one GPU. Average inference run time for UNICMR$_{\text{large}}$ is approximately 10 minutes on dev set and 21 minutes on test set with one GPU.

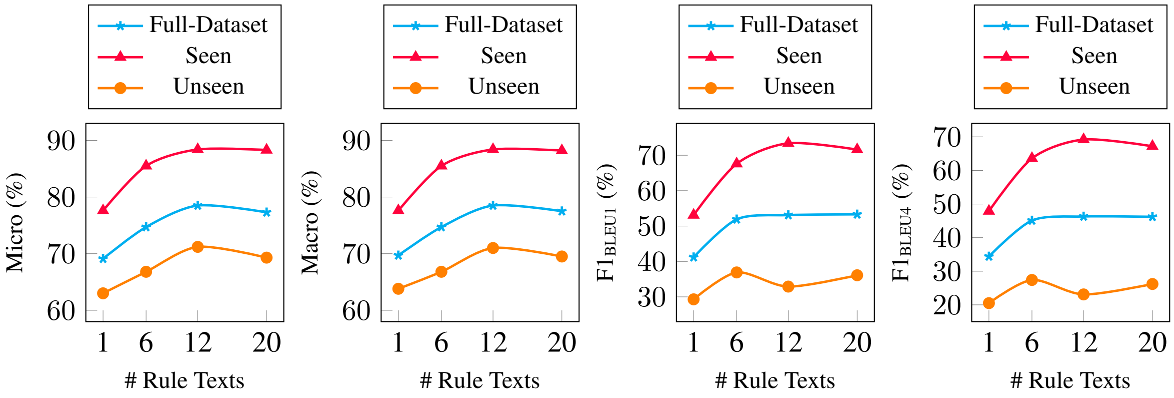


Figure 3: Evaluation performance of our model under different number of retrieved rule texts on test set.

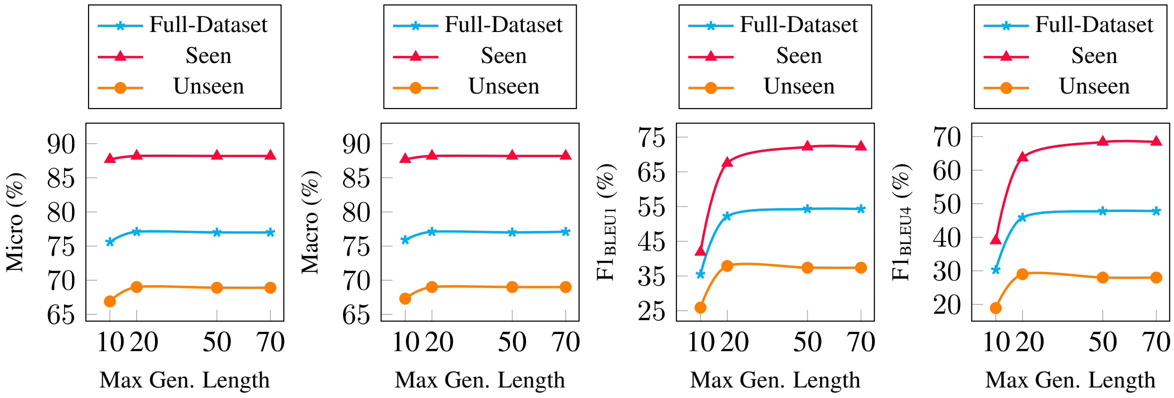


Figure 4: Evaluation performance of our model under different max generation length on test set.

beyond OR-ShARC.

	DISCERN	OSCAR	UNICMR (base/large)
#Param.	330M	1100M	220M/770M

Table 3: The comparison of approximate number of parameters of some current models.

5.5 Comparison of Model Parameter Numbers.

We have approximated total parameters of current high performance models. The information is shown in Table 3. By comparison of the parameter numbers used in current high performance models in Table 3, our UNICMR$_{\texttt{large}}$ (based on T5-large) uses around 770M parameters which generally prevails the current SOTA model OSCAR using around 1100M parameters. Our UNICMR$_{\texttt{base}}$ (based on T5-base) uses 220M parameters but prevails models like DISCERN which uses around 330M parameters. UNICMR$_{\texttt{base}}$ also achieves a close performance to OSCAR in terms of question generation. The above observations verify that our method of unifying optimizing the two sub-tasks is effective, which enables each sub-task to benefit from the optimization of the other task.

6 Analysis

6.1 Number of Retrieved Rule Texts

The model performance under different choices of the number of retrieved rule texts is shown in Table 7 in Appendix B whose visualization is shown by Figure 3. We see that generally, when the number of rule texts increases, there will be more information which improves our model while also introducing more noise which harms our model. In terms of decision making, our model is quite stable in seen test dataset when the number of rule texts varies. That means our model well captures the useful and trash conditions in rule texts and fulfillment states in dialogue history in the training stage. Besides, The unusual boost of question generation performance in the unseen test set might suggest that using more than the necessary number of rule texts possibly pushes the model to gain more power of generalization in the training stage.

Model	Dev Set				Test Set			
	Micro	Macro	F1$_{\text{BLEU1}}$	F1$_{\text{BLEU4}}$	Micro	Macro	F1$_{\text{BLEU1}}$	F1$_{\text{BLEU4}}$
UNICMR$_{\texttt{large}}$	77.7	78.0	59.3	52.8	76.7	76.7	54.2	47.9
Closed-Book	**82.1**	**82.1**	**67.8**	**62.8**	**79.4**	**79.5**	**60.5**	**54.8**
w/ DPR++	76.8	77.4	56.8	50.4	75.2	75.2	54.8	48.8
w/o Retriever	71.0	70.9	42.1	35.2	65.8	65.7	35.2	28.7

Table 4: Results of our UNICMR$_{\texttt{large}}$ and UNICMR$_{\texttt{large}}$ with different retriever module setting on the dev and test sets of OR-ShARC benchmark. For Closed-Book setting, the OR-ShARC is turned into a closed-book setting by given the rule texts. For w/ DPR++ setting, the TF-IDF retriever is replaced with DPR++ retriever. For w/o Retriever setting, the OR-ShARC is approached without rule texts.

6.2 Maximum Generation Length

The model performance under the different choices of the maximum generation length is shown in Table 8 [6] in Appendix B whose visualization is shown by Figure 4.

In terms of decision making and question generation, redundant max generation length will not affect the performance of the model but insufficient max generation length will limit the model performance. This means the model well learns the difference between different forms of answers and is able to generate answers of suitable length accordingly. This verifies the feasibility of our end-to-end framework design.

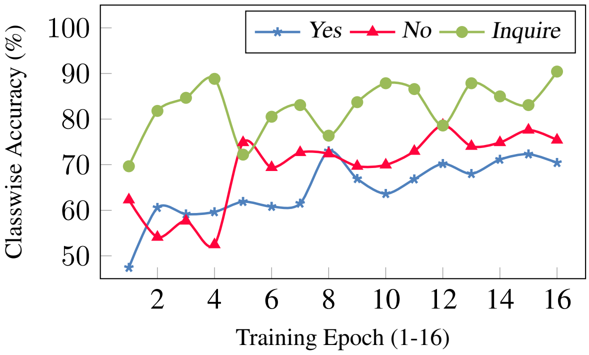


Figure 5: Classwise accuracy on dev set of each epoh.

6.3 Generation Quality Gain Across Training

The classwise accuracy evaluated in the training of the decision making sub-task is shown by Figure 5. By the initial gap between the accuracy for "Inquire" and the accuracy for other decisions, our model tends to predict the decision as *Inquire* and generate question when not well fine-tuned. This is due to a gap between the length for the answer *Yes*/*No* and the length for the answer "*Inquire*+Generated Question". And also the innate property of pre-trained T5 generation model before well fine-tuned at the beginning which is hence biased towards the longer answer. As the training continues, the accuracy for *Yes* and *No* gradually catches up with *Inquire* even though is slightly lower. This observation shows the existence of the bias of our backbone model and also the effectiveness of our training which large reduces such bias. This also suggests future improvements on more targeted training to eliminate the bias and lessening the discontinuity between the length of output for *Yes*/*No* and the length of output for "*Inquire*+Generated Question".

6.4 Contribution of the Retriever Module

To quantify the contribution of the retriever module, we conducted an additional experiment where OR-ShARC is turned into a closed-book setting (see Closed-Book in Table 4). Also, we replaced the TF-IDF retriever with the DPR++ retriever introduced in UNICMR$_{\texttt{large}}$ for reference (see w/ DPR++ in Table 4). Performance of UNICMR$_{\texttt{large}}$ without retriever is also shown (see w/o Retriever in Table 4). The results verify that using the retrieval is beneficial, which reduces the gap between the challenging open-retrieval task and the closed-book task with gold rule texts.

6.5 Discussions of Performance Improvement

To further investigate the source of performance improvement of our method, more comprehensive experimental results are shown here following the deduced conclusions.

First, UNICMR's unified training format advances the performance of training T5 separately

[6] In Table 7, the hyperparameter m (number of retrieved rule texts) is varied to compare our model performance on the OR-ShARC test set, test set seen and test set unseen divisions respectively. In Table 8, the hyperparameter maximum generation length of the backbone encoder-decoder model is varied to compare our model performance on the same datasets. The corresponding performance of the above two experiments on dev set is shown by Table 9 and Table 10 in Appendix B for reference. Note in these experiments, all the hyperparameters remain the same unless explicitly stated.

Model	Dev Set				Test Set			
	BLEU1	BLEU4	$F1_{\text{BLEU1}}$	$F1_{\text{BLEU4}}$	BLEU1	BLEU4	$F1_{\text{BLEU1}}$	$F1_{\text{BLEU4}}$
w/ T5-large								
UNICMR	67.5	59.1	59.3	52.8	55.8	48.3	54.2	47.9
QG-only whole-evaluation	53.3	47.5	49.5	43.1	45.2	40.1	47.0	39.7
QG-only partial-evaluation	71.1	61.0	47.9	40.8	69.4	59.5	45.8	38.9
w/ BART-base								
UNICMR	58.4	50.2	52.3	45.1	47.3	40.2	46.9	39.8
QG-only whole-evaluation	62.6	51.4	44.1	37.3	60.4	48.9	40.3	33.5
QG-only partial-evaluation	69.2	57.7	43.3	35.3	66.8	56.7	39.9	33.3

Table 5: Question generation performance of UNICMR compared with models trained only on question generation sub-task on OR-ShARC. For QG-only whole-evaluation setting, we use all samples by assigning empty generated question to samples with *Yes*/*No* decisions. For QG-only partial-evaluation setting, we use samples only with inquiry questions. The results are generally divided into two parts, one using T5-large as backbone model and one using BART-base as backbone model.

Model	Dev Set		Test Set	
	Micro	Macro	Micro	Macro
w/ T5-large				
UNICMR	77.7	78.0	76.7	76.7
DM-only	73.9	73.7	72.9	72.3
w/ BART-base				
UNICMR	74.8	75.7	71.5	71.8
DM-only	72.5	72.4	68.6	68.3

Table 6: Decision making performance of UNICMR compared with models trained only on decision making sub-task on OR-ShARC. For DM-only setting, we use all samples to train our model only on decision making sub-task. The results are generally divided into two parts, one using T5-large as backbone model and one using BART-base as backbone model.

on decision making. See the performance of T5-large trained for decision making separately (DM-only in Table 6) compared with the original UNICMR$_{\texttt{large}}$ (UNICMR in Table 6) performance. The comparison indicates that UNICMR's stronger form of unified training improves the model's decision making ability.

Second, UNICMR's unified training format advances the performance of training T5 separately on question generation in $F1_{\text{BLEU}}$. Ablation studies here include the T5-large trained with all examples (assign empty to examples with *Yes* and *No* decisions) for question generation only (QG-only whole-evaluation in Table 5), T5-large trained with examples with gold inquiry questions for questions generation only (QG-only partial-evaluation in Table 5), and T5-large-based UNICMR (UNICMR in Table 5). The results indicate that:

(i) In terms of $F1_{\text{BLEU}}$, UNICMR has dominantly higher performance than other separately trained models.

(ii) In terms of BLEU[7], UNICMR is not the best, which shows its source of $F1_{\text{BLEU}}$ dominance includes reduction of error propagation.

(iii) For T5-large backbone, UNICMR is higher in BLEU than QG-only partial-evaluation, which means UNICMR's integration of decision making labels in training is effective.

6.6 Generalizability on Different Backbone Models

Replacing the T5-large backbone with BART-base, and repeating the same experiments (see the same settings but with BART-base as backbone models in Table 6 and Table 5), leads to same general conclusions. This shows the effectiveness of UNICMR's unified format can well generalize to different end-to-end architectures.

6.7 Error Analysis and Case Study

To reveal more insights into UNICMR, we randomly collect test samples and conduct error analysis (see Figure 6) and case study (see Figure 7 in Appendix A). The ground truth answers are indicated in red, the TF-IDF scores are indicated in green, and the predictions of UNICMR$_{\texttt{large}}$ are indicated in blue. The retrieved rule texts are in descending order in terms of TF-IDF scores.

Error Analysis. The observed test errors are summarized into four aspects: (1) *Noisy Retrieved Rule Texts* which is caused by the innate deficiencies of TF-IDF retriever with bigram

[7] Note that BLEU is measured on samples with *Inquire* as gold labels only while $F1_{\text{BLEU}}$ is measured on all samples considering both the BLEU when prediction is *Inquire* and the BLEU when gold label is *Inquire*. For $F1_{\text{BLEU}}$ calculation of all QG-only settings, decision making predictions of model trained only on decision making sub-task are used.

Error Type	*Dialogue Setups*	*UniCMR Output*
Noisy Retrieved Rule Texts	Scen.: *It was a donation of stuff I wasn't using that I gave to Gift Aid and they got me 25% more than anyplace else would.* Ques.: *Will I have to pay more tax than I've paid?* His.: *(empty)* Gold Rule: *Charity donations: tax relief. If the charity or CASC gets back more tax than you've paid, HMRC may ask you to pay more tax to cover the difference.* Gold Answer: *No*	Retrieved Rules: *(1) [78.43] Donations through Gift Aid: Charities and community amateur sports clubs (CASCs) can register with HM Revenue and Customs (HMRC) to be part of the Gift Aid scheme. When they're registered, they can claim back the tax you've already paid on your donation.* *(2) [39.25] Charity donations: tax relief. Donations to charity from individuals are tax free. You can get tax relief if you donate: * through Gift Aid * straight from your wages or pension, through Payroll Giving* …… Prediction: *Can a charity or community amateur sports club (CASC) register with HM Revenue and customs (HMRC)?*
Losing Track of Some Conditions	Scen.: *I married my husband Bob when he was 50 in the year 2014. Unfortunately he died of a heart attack in 2015 and never reached the age where he was eligible for a state pension.* Ques.: *Could I inherit part of my deceased partner's Additional State Pension?* His.: *Q: Did you partner reach state pension age before April 6 2016? A: No* Gold Rule: *(same as the first retrieved rule)* Gold Answer: *Would your partner have reached state pension age on or after 6 April 2016?*	Retrieved Rules: *(1) [149.46] Inheriting Additional State Pension. You might inherit part of your deceased partner's Additional State Pension if your marriage or civil partnership with them began before 6 April 2016 and one of the following applies: * your partner reached State Pension age before 6 April 2016 * they died before 6 April 2016 but would have reached State Pension age on or after that date* *(2) [99.16] Inheriting or increasing State Pension from a spouse or civil partner. You may inherit part of or all of your partner's extra State Pension or lump sum if: * they died while they were deferring their State Pension (before claiming) or they had started claiming it after deferring * they reached State Pension age before 6 April 2016 * you were married or in the civil partnership when they died* …… Prediction: *Did they die before April 6 2016?*
Different Condition Ordering	Scen.: *I require assistance in turning a local agricultural area into a hunting and gaming area.* Ques.: *Do I get the Additional State Pension automatically?* His.: *(empty)* Gold Rule: (same as the first retrieved rule) Gold Answer: *Have you contracted out of the state?*	Retrieved Rules: *(1) [49.66] Overview. You get the Additional State Pension automatically if you're eligible for it, unless you've contracted out of it.* *(2) [41.62] Inheriting Additional State Pension. You might inherit part of your deceased partner's Additional State Pension if your marriage or civil partnership with them began before 6 April 2016 and one of the following applies: * your partner reached State Pension age before 6 April 2016 * they died before 6 April 2016 but would have reached State Pension age on or after that date* …… Prediction: *Are you eligible for it?*
BLEU's Inability on Phrase Variants	Scen.: *(empty)* Ques.: *Can I get payment in lieu?* His.: *(empty)* Gold Rule: (same as the first retrieved rule) Gold Answer: *Are you leaving your job?*	Retrieved Rules: *(1) [21.83] Getting paid instead of taking holidays. The only time someone can get paid in place of taking statutory leave (known as 'payment in lieu') is when they leave their job. Employers must pay for untaken statutory leave (even if the worker is dismissed for gross misconduct).* …… Prediction: *Did you leave your job?*

Figure 6: Error analysis of UniCMR$_{\texttt{large}}$ by comparison with ground truth answers.

features. (2) *Losing Track of Some Conditions* which shows in rare cases UniCMR$_{\texttt{large}}$ might miss some condition fulfillment as UniCMR$_{\texttt{large}}$ does not explicitly model condition fulfillment. (3) *Different Condition Ordering* which is caused by multiple unsatisfied conditions and the flexibility to inquire any of them. (4) *BLEU's Inability on Phrase Variants* which means predictions are penalized by BLEU even if they only differ in unimportant and semantically harmless words.

Case Study. Qualitative improvements of generated inquiries of UniCMR$_{\texttt{large}}$ are summarized into two aspects: (1) *Exactness* which means the capability of capturing the self-contained yet elementary condition units that need to be clarified. (2) *Robustness to Noisy Retrieved Rules* which means the model can filter noisy retrieved rule texts to extract unsatisfied conditions. From the results in Figure 7, it can be seen that UniCMR$_{\texttt{large}}$ generate more suitable inquires in terms of *Exactness* and achieves excellent performance in terms of *Robustness to Noisy Retrieved Rules*. This suggests that our fully end-to-end framework enables the accurate focus on target conditions and the implicit feature engineering of UniCMR is powerful to filter noisy retrievals regardless of the retriever quality.

7 Conclusion

In this paper, we study open-retrieval setting of the conversation machine reading task and promote a novel framework to first unify the optimizations of the two sub-tasks to achieve optimization synergy. With a retriever module and a parameter-efficient text-to-text encoder-decoder module, we have achieved new SOTA results in both the CMR and the OR-CMR benchmarks. Further experiments shows that our unified training form with an end-to-end optimization method largely contributes to the advanced performance in decision making and reduces the error propagation to boost question generation performance. It's also shown that our framework well generalize to other backbone models. Further qualitative analysis also verifies our framework's effectiveness.

Limitations

Under the challenging open-retrieval setting, a retrieval is required to find the related rules texts. However, the performance of our model may be hindered by the noise introduced by the irrelevant rule texts from the retrieval. To conquer this deficiency, it is beneficial to develop additional filtering methods to alleviate the influence of irrelevant rule texts.

References

Wei Bi, Jun Gao, Xiaojiang Liu, and Shuming Shi. 2019. Fine-grained sentence functions for short-text conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3984–3993, Florence, Italy. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau 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*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Yifan Fan, Xudong Luo, and Pingping Lin. 2020. A survey of response generation of dialogue systems. *International Journal of Computer and Information Engineering*, 14(12):461–472.

Yifan Gao, Jingjing Li, Michael R Lyu, and Irwin King. 2021. Open-retrieval conversational machine reading. *arXiv preprint arXiv:2102.08633*.

Yifan Gao, Piji Li, Wei Bi, Xiaojiang Liu, Michael R. Lyu, and Irwin King. 2020a. Dialogue generation on infrequent sentence functions via structured meta-learning.

Yifan Gao, Chien-Sheng Wu, Shafiq Joty, Caiming Xiong, Richard Socher, Irwin King, Michael Lyu, and Steven C.H. Hoi. 2020b. Explicit memory tracker with coarse-to-fine reasoning for conversational machine reading. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 935–945, Online. Association for Computational Linguistics.

Yifan Gao, Chien-Sheng Wu, Jingjing Li, Shafiq Joty, Steven C.H. Hoi, Caiming Xiong, Irwin King, and Michael Lyu. 2020c. Discern: Discourse-aware entailment reasoning network for conversational machine reading. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2439–2449, Online. Association for Computational Linguistics.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3682–3692, Online. Association for Computational Linguistics.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *International Conference on Learning Representations*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Carolin Lawrence, Bhushan Kotnis, and Mathias Niepert. 2019. Attending to future tokens for bidirectional sequence generation. 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 1–10, Hong Kong, China. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Dan Moldovan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Roxana Girju, Richard Goodrum, and Vasile Rus. 2000. The structure and performance of an open-domain question answering system. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 563–570, Hong Kong. Association for Computational Linguistics.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2021. Dialogue graph modeling for conversational machine reading. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Da Ren, Yi Cai, Xue Lei, Jingyun Xu, Qing Li, and Ho fung Leung. 2019. A multi-encoder neural conversation model. *Neurocomputing*, 358:344–354.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7007–7014. AAAI Press.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 267–275, New York, NY, USA. Association for Computing Machinery.

Nikhil Verma, Abhishek Sharma, Dhiraj Madan, Danish Contractor, Harshit Kumar, and Sachindra Joshi. 2020. Neural conversational QA: Learning to reason vs exploiting patterns. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7263–7269, Online. Association for Computational Linguistics.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Bowen Wu, Mengyuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2019. Guiding variational response generator to exploit persona. In *Annual Meeting of the Association for Computational Linguistics*.

Liu Yang, Minghui Qiu, Chen Qu, Cen Chen, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, and Haiqing Chen. 2020. Iart: Intent-aware response ranking with transformers in information-seeking conversation systems. In *Proceedings of The Web Conference 2020*, WWW '20, page 2592–2598, New York, NY, USA. Association for Computing Machinery.

Munazza Zaib, Quan Z Sheng, and Wei Emma Zhang. 2020. A short survey of pre-trained language models for conversational ai-a new age in nlp. In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–4.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zhuosheng Zhang, Siru Ouyang, Hai Zhao, Masao Utiyama, and Eiichiro Sumita. 2021. Smoothing dialogue states for open conversational machine reading. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3685–3696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Zhong and Luke Zettlemoyer. 2019. E3: Entailment-driven extracting and editing for conversational machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2310–2320, Florence, Italy. Association for Computational Linguistics.

Pengfei Zhu, Zhuosheng Zhang, Jiangtong Li, Yafang Huang, and Hai Zhao. 2018. Lingke: a fine-grained multi-turn chatbot for customer service. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 108–112, Santa Fe, New Mexico. Association for Computational Linguistics.

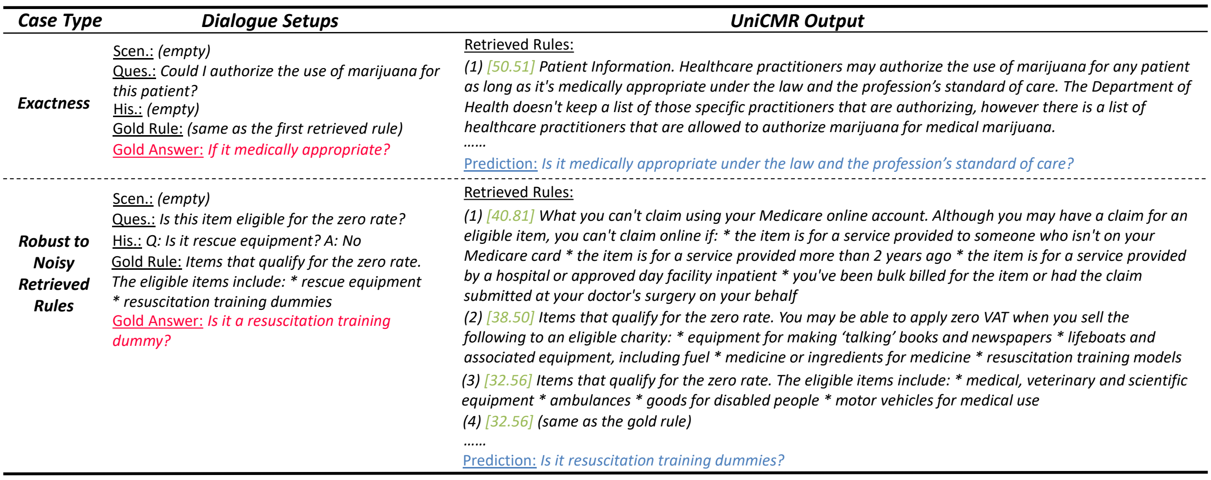
A Case Study

To reveal more insights into our framework, we randomly collect test samples and conduct the case study (see Figure 7). The ground truth answers are indicated in red, the TF-IDF scores are indicated in green, and the predictions of UNICMR$_{\texttt{large}}$ are indicated in blue. The retrieved rule texts are in descending order in terms of TF-IDF scores. For the analysis on cases, please refer to Section 6.7.

B Hyperparameter-Related Experiments

In this section, additional experiments related to the hyperparameter m (number of retrieved rule texts) and the hyperparameter maximum generation length are conducted with their results shown in Table 7-10.

In Table 7, the hyperparameter m (number of retrieved rule texts) is varied to compare our model performance on the OR-ShARC test set, test set seen and test set unseen divisions respectively. In Table 8, the hyperparameter maximum generation length of the backbone encoder-decoder model is varied to compare our model performance on the same datasets. The corresponding performance of the above two experiments on dev set is shown by Table 9 and Table 10 respectively. Note in these experiments, all the hyperparameters remain the same unless explicitly stated.



Case Type	*Dialogue Setups*	*UniCMR Output*
Exactness	Scen.: *(empty)* Ques.: *Could I authorize the use of marijuana for this patient?* His.: *(empty)* Gold Rule: *(same as the first retrieved rule)* Gold Answer: *If it medically appropriate?*	Retrieved Rules: *(1) [50.51] Patient Information. Healthcare practitioners may authorize the use of marijuana for any patient as long as it's medically appropriate under the law and the profession's standard of care. The Department of Health doesn't keep a list of those specific practitioners that are authorizing, however there is a list of healthcare practitioners that are allowed to authorize marijuana for medical marijuana.* …… Prediction: *Is it medically appropriate under the law and the profession's standard of care?*
Robust to Noisy Retrieved Rules	Scen.: *(empty)* Ques.: *Is this item eligible for the zero rate?* His.: *Q: Is it rescue equipment? A: No* Gold Rule: *Items that qualify for the zero rate. The eligible items include: * rescue equipment * resuscitation training dummies* Gold Answer: *Is it a resuscitation training dummy?*	Retrieved Rules: *(1) [40.81] What you can't claim using your Medicare online account. Although you may have a claim for an eligible item, you can't claim online if: * the item is for a service provided to someone who isn't on your Medicare card * the item is for a service provided more than 2 years ago * the item is for a service provided by a hospital or approved day facility inpatient * you've been bulk billed for the item or had the claim submitted at your doctor's surgery on your behalf* *(2) [38.50] Items that qualify for the zero rate. You may be able to apply zero VAT when you sell the following to an eligible charity: * equipment for making 'talking' books and newspapers * lifeboats and associated equipment, including fuel * medicine or ingredients for medicine * resuscitation training models* *(3) [32.56] Items that qualify for the zero rate. The eligible items include: * medical, veterinary and scientific equipment * ambulances * goods for disabled people * motor vehicles for medical use* *(4) [32.56] (same as the gold rule)* …… Prediction: *Is it resuscitation training dummies?*

Figure 7: Case study of UNICMR$_{\texttt{large}}$ by comparison with ground truth answers.

Test Set	Decision Making								Question Generation							
	Micro				Macro				F1$_{\text{BLEU1}}$				F1$_{\text{BLEU4}}$			
	1	6	12	20	1	6	12	20	1	6	12	20	1	6	12	20
Full-Dataset	69.1	74.7	78.5	77.3	69.7	74.7	78.5	77.5	41.2	51.9	53.1	53.3	34.4	45.1	46.3	46.2
Seen	77.6	85.5	88.4	88.3	77.6	85.5	88.4	88.2	53.1	67.6	73.4	71.6	47.9	63.6	69.2	67.2
Unseen	63.0	66.8	71.2	69.3	63.8	66.8	71.0	69.5	29.3	36.9	32.9	36.1	20.5	27.4	23.1	26.2

Table 7: Comparison of our model under different number of retrieved rule texts on test set.

Test Set	Decision Making								Question Generation							
	Micro				Macro				F1$_{\text{BLEU1}}$				F1$_{\text{BLEU4}}$			
	10	20	50	70	10	20	50	70	10	20	50	70	10	20	50	70
Full-Dataset	75.6	77.1	77.0	77.0	75.9	77.1	77.0	77.1	35.5	52.2	54.3	54.3	30.3	45.9	47.8	47.8
Seen	87.7	88.2	88.2	88.2	87.7	88.2	88.2	88.2	41.9	67.5	72.2	72.2	39.0	63.7	68.4	68.4
Unseen	66.9	69.0	68.9	68.9	67.3	69.0	69.0	69.0	25.9	37.9	37.4	37.4	18.9	29.0	28.0	28.0

Table 8: Comparison of our model under different max generation length limit on test set.

Dev Set	Decision Making								Question Generation							
	Micro				Macro				F1$_{\text{BLEU1}}$				F1$_{\text{BLEU4}}$			
	1	6	12	20	1	6	12	20	1	6	12	20	1	6	12	20
Full-Dataset	65.4	76.6	77.8	77.6	66.3	76.7	78.2	78.2	36.6	58.2	58.9	58.8	29.5	51.6	53.3	51.8
Seen	78.4	88.2	88.8	90.6	78.2	88.1	88.8	90.5	52.7	71.8	74.6	72.6	47.5	66.8	70.7	68.4
Unseen	54.7	66.9	68.8	66.8	56.4	66.7	69.1	68.7	19.7	40.0	38.6	43.1	10.3	30.5	30.5	32.5

Table 9: Comparison of our model under different number of retrieved rule texts on dev set.

Dev Set	Decision Making								Question Generation							
	Micro				Macro				F1$_{\text{BLEU1}}$				F1$_{\text{BLEU4}}$			
	10	20	50	70	10	20	50	70	10	20	50	70	10	20	50	70
Full-Dataset	77.6	76.9	77.1	77.1	78.5	77.4	77.7	77.7	46.9	57.2	66.1	61.1	40.8	50.9	54.8	54.8
Seen	91.0	89.4	89.8	89.8	90.9	89.3	89.7	89.7	47.7	71.0	78.1	78.1	44.1	66.6	73.9	73.9
Unseen	66.5	66.6	66.6	66.6	68.5	67.7	67.7	67.7	36.3	40.6	40.7	40.1	27.6	32.2	31.6	31.6

Table 10: Comparison of our model under different max generation length limit on dev set.