Bridging The Gap: Entailment Fused-T5 for Open-retrieval Conversational Machine Reading Comprehension

Xiao Zhang[123], **Heyan Huang**[123]*, **Zewen Chi**[123], **Xian-Ling Mao**[123]
[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
[3]Southeast Academy of Information Technology, Beijing Institute of Technology
{yotta,hhy63,czw,maoxl}@bit.edu.cn

Abstract

Open-retrieval conversational machine reading comprehension (OCMRC) simulates real-life conversational interaction scenes. Machines are required to make a decision of `Yes/No/Inquire` or generate a follow-up question when the decision is `Inquire` based on retrieved rule texts, user scenario, user question and dialogue history. Recent studies try to reduce the information gap between decision-making and question generation, in order to improve the performance of generation. However, the information gap still persists because these methods are still limited in pipeline framework, where decision-making and question generation are performed separately, making it hard to share the entailment reasoning used in decision-making across all stages. To tackle the above problem, we propose a novel one-stage end-to-end framework, called Entailment Fused-T5 (EFT), to bridge the information gap between decision-making and question generation in a global understanding manner. The extensive experimental results demonstrate that our proposed framework achieves new state-of-the-art performance on the OR-ShARC benchmark. Our model and code are publicly available[1].

1 Introduction

Open-retrieval conversational machine reading comprehension (OCMRC) (Gao et al., 2021) investigates real-life scenes, aiming to formulate multi-turn interactions between humans and machines in open-retrieval settings. As shown in Figure 1, given the user scenario and user question, machines are required to first retrieve related rule texts in the knowledge database, and then make a decision of `Yes/No/Inquire` or generate a follow-up question when the decision is `Inquire` based on retrieved rule texts, user scenario, user question and dialogue history.

Retrieved Rule Text 1: The airport manager, or his authorized…
Retrieved Rule Text 2: When to contact the benefit office**:** It's your responsibility to contact the office straight away if: You think you've been overpaid . You get a letter telling you that you have been overpaid - and who to contact your circumstances change and this is likely to affect a benefit .
Retrieved Rule Text 3: Your obligations for using center pay…
…

User Question: Is it my responsibility to contact the office in this situation
User Scenario: : I most assuredly believe…
Follow-up Q: Do you believe that you have been overpaid?
Follow-up A : No
Follow-up Q: Did you receive a letter stating that you were overpaid?
Follow-up A : No

Decision-Making: Yes | No| **Inquire**
Final Answer: **Have your circumstances changed ?**

Figure 1: An example in the OCMRC dataset. Given the user scenario and user question, machines are required to first retrieve related rule texts in the knowledge database, and then make a decision of `Yes/No/Inquire` or generate a follow-up question when the decision is `Inquire` based on retrieved rule texts, user scenario, user question and dialogue history.

Previous studies (Saeidi et al., 2018; Verma et al., 2020; Lawrence et al., 2019; Zhong and Zettlemoyer, 2019; Gao et al., 2020a,b; Ouyang et al., 2021; Gao et al., 2021; Zhang et al., 2021) typically adopt pipeline frameworks based on pre-trained language models (PrLM) (Devlin et al., 2019; Clark et al., 2020; Lewis et al., 2020; Liu et al., 2020) , as shown in Figure 4, these frameworks usually consist of three stages, includ-

*Corresponding author.
[1]github.com/Yottaxx/EFT

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15374–15386

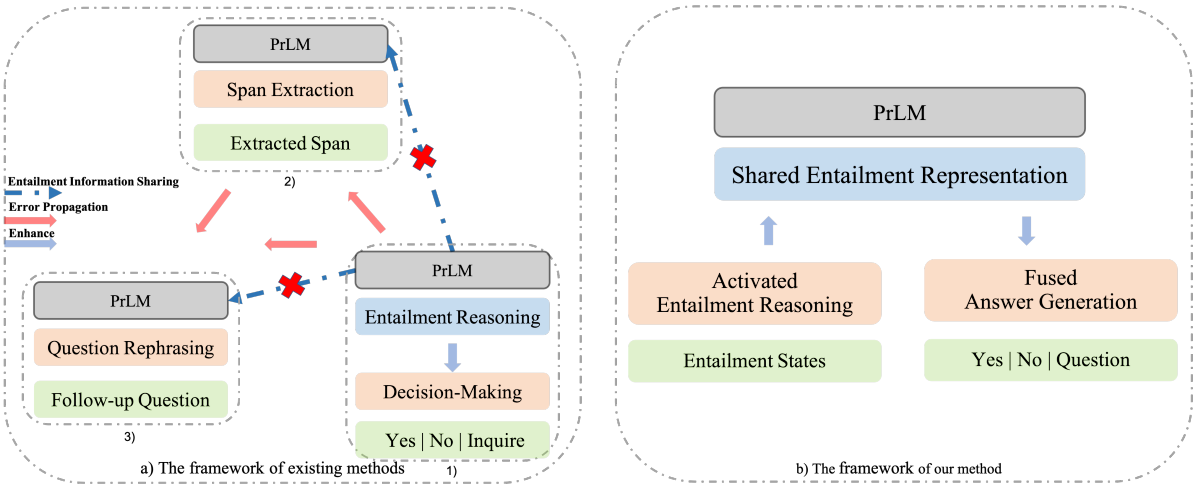

Figure 2: The comparison between our framework and previous pipeline framework. a) Previous framework typically has three stages: entailment reasoning based decision-making, span extraction and question rephrasing. Thus, the entailment reasoning utilized in decision-making is hard to share through all stages. Meanwhile, the performance of previous framework suffers from error propagation problem. The above information misleading leads to the information gap between decision and generation. b) Our framework is a one-stage end-to-end model. To bridge the information gap, the fused answer generation directly generates the decisions or follow-up questions with the shared entailment representation enhanced by activated entailment reasoning.

ing decision-making, span extraction and question rephrasing. Different entailment reasoning strategies are utilized to improve the performance of decision-making. Span extraction and question rephrasing are conducted for question generation. These pipeline frameworks are either completely independent of the three stages (Zhong and Zettlemoyer, 2019; Gao et al., 2020a,b; Ouyang et al., 2021; Gao et al., 2021), or try to reduce the information gap between decision-making and question generation through representation-fused methods (Zhang et al., 2021) among three stages .

However, the information gap still persists because these methods are still limited in pipeline framework, where decision-making and question generation are performed separately, making it hard to share the entailment reasoning used in decision-making across all stages.

To tackle the above problem, we propose a novel one-stage end-to-end framework, called entailment fused-T5 (EFT) to bridge the information gap between decision-making and question generation in a global understanding manner. Specifically, our model consists of a universal encoder and a duplex decoder. The decoder consists of an activated entailment reasoning decoder and a fused answer generation decoder. The implicit reasoning chains of both decision-making and question generation in the multi-fused answer generation decoder are explicitly supervised by activated entailment reasoning through the shared entailment representation of our encoder. Moreover, a relevance-diversity fusion strategy is utilized to further improve the implicit reasoning abilities among the multiple retrieved rules for the fused answer generation decoder through the implicit ranking method. Thus, our model can reason in a global understanding manner. The extensive results, as illustrated in Figure 3, demonstrate that our proposed framework EFT achieves new state-of-the-art performance on the OR-ShARC benchmark.

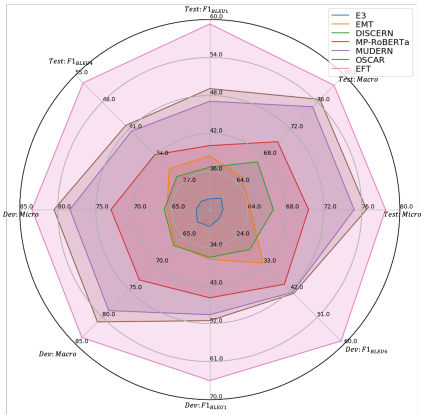

Figure 3: EFT achieves state-of-the-art performance on the OR-ShARC benchmark.

Our contributions are summarized as follows:

- We propose a novel one-stage end-to-

end framework, called entailment fused-T5 (EFT) to bridge the information gap between decision-making and question generation through a global understanding manner.

- We further investigate a relevance-diversity fusion strategy (RD strategy) to improve the implicit reasoning abilities of our model.

- Extensive experiments demonstrate the effectiveness of our proposed framework on the OR-ShARC benchmark.

2 Related Work

Conversation-based Reading Comprehension Conversation-based reading comprehension (Saeidi et al., 2018; Sun et al., 2019; Reddy et al., 2019; Choi et al., 2018; Cui et al., 2020; Gao et al., 2021) aims to formulate human-like interactions. Compared to traditional reading comprehension, these tasks extend the reading comprehension scenarios with dialogue interactions. There are typically three main types of these tasks: span-based QA tasks (Choi et al., 2018; Reddy et al., 2019), multi-choice tasks (Sun et al., 2019; Cui et al., 2020), or hybrid-form tasks (Saeidi et al., 2018; Gao et al., 2021).

Conversational Machine Reading Comprehension CMRC (Saeidi et al., 2018) is the hybrid form of conversation-based reading comprehension, which requires the machines to make a decision or generate a follow-up question based on rule text, user scenario, user question and dialogue history.

In this paper, we focus on the open-retrieval conversational machine reading (OCMRC) task (Gao et al., 2021), which further extends the CMRC task into a real-life scenario. Machines are required to first retrieve related rule texts in a knowledge base based on user questions and user scenarios, then machines are required to make a decision of `Yes/No/Inquire`, or a follow-up question if the decision is `Inquire` based on the relevant rule texts, user scenario, user question and dialogue history.

Due to the hybrid-form task, the previous methods (Zhong and Zettlemoyer, 2019; Gao et al., 2020a,b; Ouyang et al., 2021; Zhang et al., 2021) typically adopt pipeline architectures, including decision-making, span extraction and question phrasing. Various kinds of entailment reasoning strategies are proposed to improve the performance of decision-making. Despite the effectiveness of entailment reasoning, the performance is still limited because of the information gap between decision-making and question generation. Recent studies (Zhang et al., 2021, 2022) explored entailment reasoning sharing methods to reduce the gap between decision-making and question generation, but the performance is limited due to its frame flaws. In this paper, we propose a novel one-stage end-to-end model, called entailment fused-T5 (EFT), the details are written in the next sections.

3 Methods

In open-retrieval CMRC, the machines are first required to retrieve related rule texts in a knowledge base, given user question and user scenario. Then machines are required to make decisions or generate follow-up questions based on retrieved rule texts, user scenario, user question and dialogue history. Thus, we conduct a retriever to first retrieve related rule texts from the knowledge database, and then generate the final answer through our end-to-end reader EFT. The training procedure is shown in Algorithm 1.

3.1 Retriever

We first concatenate the user question and user scenario as the query to retrieve related rule texts in the knowledge base. The knowledge base is divided into the seen subset and the unseen subset. This is to simulate the real usage scenario: users will ask questions about rules they have already seen, or rules that are completely new. We only use seen rules in the training process. In this work, we utilize DPR (Karpukhin et al., 2020) to retrieve related rule texts. In contrast to previous approaches (Zhang et al., 2021) that employ TF-IDF negatives as DRP hard negatives and restrict the scope of retrieved negatives to a limited data space, we adopt a different strategy. We randomly sample rule texts from the known knowledge base to serve as the negatives. Each step will randomly sample m numbers negatives in the training stage. We retrieve the top 20 relevant rule texts for each query, which is further used by our reader.

3.2 Reader: EFT

In this stage, each item is formed as the tuple $\{R, S, Q, D\}$. R donates the rule text can-

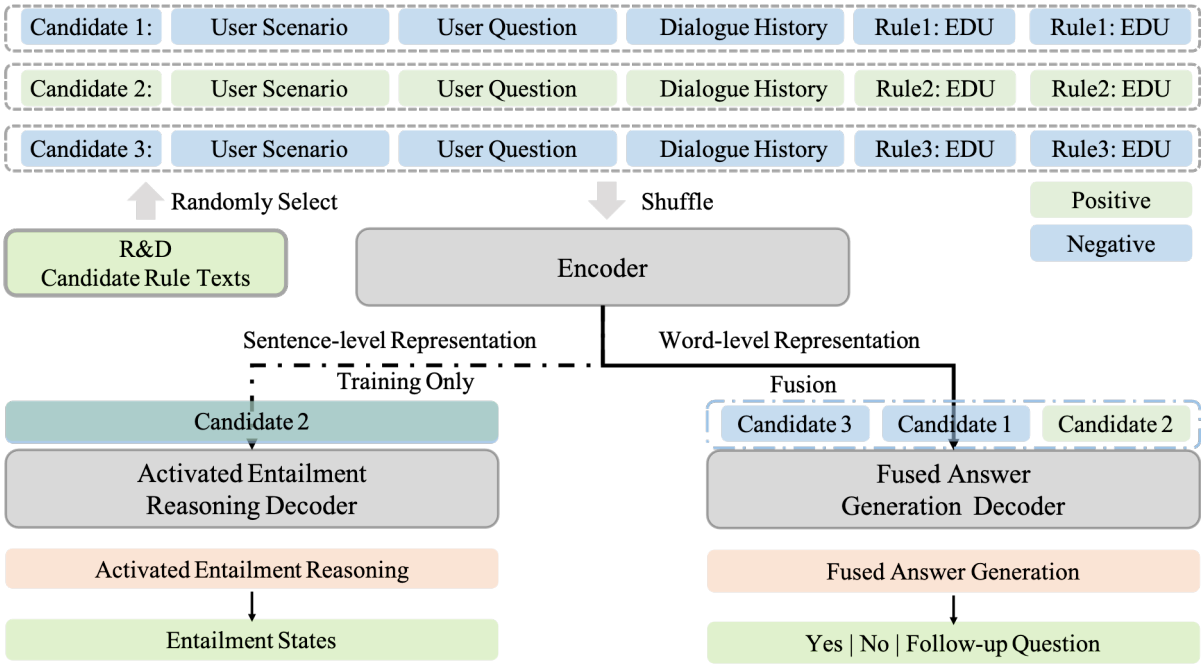

Figure 4: The architecture of our proposed EFT. Machines first randomly select related rule texts from RD candidate rule texts in the training stage, while in the evaluating stage, machines only use the top-5 retrieved rule texts. Then the input representation is encoded separately, the sentence-level representation is utilized for activated entailment reasoning, and the word-level representation is fused for the final answer generation.

didates. $R = \{r_1, r_2, ..., r_k\}$, where r donates the rule text item of R. S and Q represent user scenario and user questions, respectively. D donates the dialogue history. Given $\{R, S, Q, D\}$, EFT will directly generate a decision of `Yes/No/Inquire` or follow-up question when the decision is `Inquire`. EFT consists of a universal encoder and a duplex decoder. The Duplex decoder consists of an activated entailment reasoning decoder and a fused answer generation decoder. In this way, the whole implicit reasoning chains of the fused answer generation decoder will be fine-grained supervised by activated entailment reasoning with the shared entailment representation. Thus, the fused answer generation decoder will reason in a global understanding manner. The details are shown in Figure 4.

3.3 Encoding

Given $\{R, S, Q, D\}$, we random sample k items in R, and concatenate each of them with S, Q, D as c, thus the item collection is formed as $C = \{c_1, c_2, ..., c_k\}$. Specifically, each r in R is first parsed to elementary discourse unit (EDU) by a pre-trained model (Li et al., 2018). The final input format is shown in Figure 4. To prevent the leakage of location information in the fused answer generation decoder, and enhance the information extraction ability of the decoder, we utilize a relevance-diversity fusion (RD) strategy to randomly shuffle the order of items which are sampled from RD candidate rule texts, the details are written in Sec 3.4. Given C, we utilize T5 encoder as our backbone to get the representation. The presentation of special token are utilized as sentence-level representation $H_s = \{h_{s1}, h_{s2}, ..., h_{sk}\}$ for activated entailment reasoning decoding. The word-level representation $H_w = \{h_{w1}, h_{w2}, ..., h_{wk}\}$ are utilized for fused answer generation decoding.

3.4 Decoding

We utilize duplex decoding to explicitly supervise our answer generation stage, which introduced the explicit entailment reasoning information in implicit answer generation reasoning. The answer generation will either generate a decision of `Yes/No/Inquire` or a follow-up question when the decision is `Inquire`. The activated entailment reasoning decoder will reason the entailment states of the EDUs. The duplex decoder is trained in a multi-task form. And the activated entailment reasoning only activates in training stage.

Activated Entailment Reasoning Each EDU will be classified into one of three entailment states, including ENTAILMENT, CONTRADICTION and NEUTRAL. To get the noisy supervision signals of entailment states, we adopt a heuristic approach[2]. This is proposed to simulate fulfillment prediction of conditions in multi-sentence entailment reasoning, which can explicitly supervise the implicit reasoning chains of the answer generation.

Previous studies typically introduce entailment reasoning in all rule text segmented EDUs. This will greatly increase the proportion of NEUTRAL labels and affect the model effect, because nearly all of the entailment states of EDUs in retrieved irrelevant rules are NEUTRAL, and introducing more noise in the training stage. In our method, entailment reasoning will only activate for the golden rule text. Utilizing this setting, one benefit is to balance the categories of entailment reasoning, and the other is to supervise the implicit reasoning of the fused decoder, which can help the fused decoder infer correct rule text from multiple retrieved rule texts.

Given the sentence-level representation H_s, we utilize inter-attention reasoning to fully interact with various information in r, including EDUs, user question, user scenario and dialogue history. We utilize an inter-sentence Transformer (Devlin et al., 2019; Vaswani et al., 2017) to get the interacted sentence-level representation G_s. Then, we use a linear transformation to track the entailment reasoning states of each EDU in activated rule text.

$$e_i = W_c \tilde{h}_{e_i} + b_c \in \mathcal{R}^3, \tag{1}$$

where the W_c is trainable parameters, e_i is the predicted score for the three labels of the i-th states.

Fused Answer Generation Given the word-level representation $H_w = \{h_{w1}, h_{w2}, ..., h_{wk}\}$ of R, we concatenate the H_w as the fused representation f_w. In this manner, our answer generation decoder can reason through all the k items through an implicit ranking mechanism. It is worth mentioning that each item of H_w is first fully interacted among rule text, user question, user scenarios and dialogue history without other multi-rule noise through our encoder.

To improve the information implicit reasoning abilities of our model, we further investigate the relevance-diversity fusion strategy (RD fusion strategy), which consists of relevance-diversity candidate rule texts, order information protection and fused answer generation. The rule text candidates are consists of top k relevant rule texts and randomly sampled rule texts, which are called RD candidate rule texts. Thus, the candidates are full-filled with relevant and diverse rule texts. On the premise of ensuring relevance among the rule texts, the diversity of learnable information sampling combinations is further improved. Moreover, the order of items fused in f_w may lead to information leakage and affect the reasoning ability of the decoder in the training stage, so as we mentioned in the last section, we will randomly shuffle the order of items when inputting to the encoder to protect the order information. In the evaluation stage, only the top 5 unshuffled relevant rule texts will be utilized for answer generation.

The fused answer generation is utilized to generate either the decision or the follow-up question. We employ T5 decoder as our answer generation decoder. Given encoder fused representation f_w, and the final answer a, including decision or follow-up question, the answer is composed of the variable-length tokens x_i, the probabilities over the tokens are shown in the blow:

$$p(a) = \prod_{1}^{m} p(x_i | x_{<i}, f_e; \theta), \tag{2}$$

Algorithm 1 Training procedure of EFT

Input: Contextualized context C, learning rate τ,
Output: Final answer A, activated entailment reasoning state E, EFT encoder parameters $\boldsymbol{\theta}_e$, EFT fused answer generation decoder parameters $\boldsymbol{\theta}_a$, EFT activated entailment reasoning decoder parameters $\boldsymbol{\theta}_d$

Initialize $\boldsymbol{\theta}_e$,$\boldsymbol{\theta}_a$,$\boldsymbol{\theta}_d$
while not converged **do**
  for $i = 1, 2, \ldots, N$ **do**
    $\boldsymbol{h}_{si}, \boldsymbol{h}_{wi} = f(c_i, \boldsymbol{\theta}_e)$ s.t. $\forall c \in \mathcal{C}$
    $\boldsymbol{e}_i = f(h_{si}, \boldsymbol{\theta}_d)$
    $\boldsymbol{a}_i = f(h_{wi}, \boldsymbol{\theta}_a)$
  end for
  $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{\theta}} \mathcal{L}$
  $\boldsymbol{\theta}_e \leftarrow \boldsymbol{\theta}_e - \tau \boldsymbol{g}$
  $\boldsymbol{\theta}_d \leftarrow \boldsymbol{\theta}_d - \tau \boldsymbol{g}$
  $\boldsymbol{\theta}_a \leftarrow \boldsymbol{\theta}_a - \tau \boldsymbol{g}$
end while

[2]The noisy supervision signal is a heuristic label obtained by the minimum edit distance.

where θ donates the trainable parameters of our decoder.

3.5 Training Objective

Activated Entailment Reasoning The activated entailment reasoning is supervised by cross-entropy loss, by given the entailment stages c_i :

$$\mathcal{L}_{enatil} = -\frac{1}{N}\sum_{i=1}^{N} log\,\text{softmax}(c_i)_r, \quad (3)$$

where r is the ground truth of entailment state.

Fused Answer Generation The fused answer generation training objective is computed as illustrated in below:

$$\mathcal{L}_{answer} = -\sum_{i=1}^{M} log\, p(x_i|x_{<i}, f_w; \theta), \quad (4)$$

The overall loss function is:

$$\mathcal{L} = \mathcal{L}_{answer} + \lambda\mathcal{L}_{entail}. \quad (5)$$

4 Experiment and Analysis

4.1 Data

Dataset The experiment dataset is OR-ShARC (Gao et al., 2021), the current OCMRC benchmark. The corpus is crawled from the government website. There is a total of 651 rule texts collected in the knowledge base. For the validation and test set, the golden rule texts are split into unseen or seen. This is to simulate the real usage scenario: users will ask questions about rules they have already seen, or rules that are completely new. The train, dev and test size is 17,936, 1,105 and 2,373, respectively. Each item consists of utterance id, tree id, golden rule document id, user question, user scenario, dialog history, evidence and the decision.

4.2 Setup

Evaluation The evaluation consists of two parts: decision-making and question generation. We utilize Micro-Acc and Macro-Acc for the results of decision-making, and use F1_{BLEU} (Gao et al., 2021) for question generation. The F1_{BLEU} is conducted to evaluate the question generation performance when the predicted decision is `Inquire`.

$$precision_{\text{BLEU}} = \frac{\sum_{i=0}^{M} \text{BLEU}(y_i, \hat{y}_i)}{M}, \quad (6)$$

Where M is the total number of `Inquire` decisions made by our model. y_i is the predicted question, $\hat{y}_i$ is the corresponding ground truth prediction. The recall of BLEU is computed in a similar way.

$$recall_{\text{BLEU}} = \frac{\sum_{i=0}^{N} \text{BLEU}(y_i, \hat{y}_i)}{N}, \quad (7)$$

where N is the total number of `Inquire` decision from the ground truth annotation,

The calculation of F1_{BLEU} is shown below:

$$\text{F1}_{\text{BLEU}} = \frac{2 \times precision_{\text{BLEU}} \times recall_{\text{BLEU}}}{precision_{\text{BLEU}} + recall_{\text{BLEU}}}. \quad (8)$$

Implementation Details We utilize the T5-base (Raffel et al., 2020) as our reader backbone, and additionally add an activated entailment reasoning decoder, whose parameters are randomly initialized. We utilize BERT (Devlin et al., 2019) as our retriever backbone , whose parameters are initialized from DPR (Karpukhin et al., 2020). For the RD strategy, we use the top-20 retrieved rule texts and 30 randomly sampled rule texts as our fused candidates in the training stage, every step we will randomly select 5 samples from the candidates. We only use seen rule texts in the knowledge base for the training stage. And we only use top 5 retrieved rule text for the inference stage. The fused number k is set as 5 for fused answer generation for both training and inference. We use AdamW (Loshchilov and Hutter, 2018) to fine-tune our model. The learning rate is hierarchically designed, the learning rate of T5 is 2e-4, and the learning rate of activated entailment decoder is 2e-5. We tried from 0.1 to 1.0 for λ, and find 0.9 is the best hyper-parameter. The beam-size is set as 5 for the answer generation.

4.3 Results

All results on the OR-ShARC benchmark are illustrated in Table 1, including dev and test set with metrics for both decision-making and question generation.

Experimental results demonstrate that our proposed methods achieve new SOTA on the OR-ShARC benchmark. EFT outperforms OSCAR by 3.6% in Micro-Acc, 3.6% in Macro-Acc for decision-making on the dev set, and outperforms OSCAR by 2.6% in Micro-Acc, 2.7% in Macro-Acc for decision-making on the test set. In particular, our proposed EFT achieves considerable

Model	Dev Set				Test Set			
	Decision Making		Question Gen.		Decision Making		Question Gen.	
	Micro	Macro	$F1_{BLEU1}$	$F1_{BLEU4}$	Micro	Macro	$F1_{BLEU1}$	$F1_{BLEU4}$
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
MP-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
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
EFT	**83.4**±0.5	**83.8**±0.5	**65.5**±1.9	**59.0**±2.0	**78.5**±0.7	**78.5**±0.7	**59.3**±0.8	**53.0**±0.8

Table 1: Results on the dev and test set of OR-ShARC. The average results with standard deviation on 5 random seeds are reported.

Model	Micro	Macro	$F1_{BLEU1}$	$F1_{BLEU4}$
EFT	**83.4**±0.5	**83.8**±0.5	**65.5**±1.9	**59.0**±2.0
-w/o s	82.9±0.6	83.4±0.5	63.8±1.6	57.0±1.8
-w/o s+a	80.7±0.8	81.1±0.9	62.4±2.3	56.3±2.3
-w/o s+a+i	80.2±0.5	80.5±0.6	61.2±1.4	55.0±1.6
-w/o s+a+i+f	71.0±1.2	71.6±0.9	49.2±0.8	43.8±0.8

Table 2: Ablation study of EFT on the dev set of OR-ShARC. The average results with standard deviation on 5 random seeds are reported.

improvement in BLEU scores. EFT outperforms OSCAR by 27.7% in $F1_{BLEU1}$, 36.9% in $F1_{BLEU4}$ for the question generation on the dev set, and outperforms OSCAR by 20.8% in $F1_{BLEU1}$, 26.5% in $F1_{BLEU4}$ for the question generation on the test set. We further to investigate the classwise accuracy performance of EFT, as shown in Table 4. Experiments show that the accuracy of each category in OR-ShARC is improved by conducting EFT framework, compared with reported baselines.

To further investigate the performance for our proposed EFT in seen and unseen settings, the performance of the split subset [3] is illustrated in Table 3. Compared with OSCAR, the seen subset performance are greatly improved through our framework EFT. EFT greatly outperforms OSCAR by 29.1% in $F1_{BLEU1}$, 36.4% in $F1_{BLEU4}$ for the question generation on the seen test set. In addition, compared with the previous pipeline architectures utilized in MURDEN and OSCAR, our model not only improves the performance, but also makes the framework of OCMRC more lightweight. We reduce the number of model parameters from 330M to 220M, which is decreased by 33.3%. The performance on the seen subset of EFT is 35.0% higher in micro-acc than seen subset, 35.3% higher in macro-acc than unseen subset. Our retrieval results are illustrated in Table 6 and Table 5. The details are illustrated in Appendix A and Appendix B, respectively.

[3]Only BLEU scores are reported in OSCAR.

4.4 Ablation Studies

The ablation studies of EFT on the dev set of OR-ShARC benchmark are shown in Table 2. There are four settings of our EFT is considered:

- **EFT-wo/s** trains the model without relevance-diversity (RD) candidate rule texts. Only top-5 randomly shuffled relevant rule texts are considered in the training stage.
- **EFT-wo/s+a** trains this model additionally remove activated entailment reasoning.
- **EFT-wo/s+a+i** trains this model further cancels random shuffle in the training stage.
- **EFT-wo/s+a+i+f** trains this model without multi rule fused answer generation, only top-1 retrieved rule text is considered.

Analysis of RD Candidate We investigate the necessity of the RD candidate rule texts. This strategy is utilized to improve the implicit reasoning abilities of our decoder by improving the learning space of fused candidates. On the premise of ensuring the relevance among the rule texts, the diversity of learnable information sampling combinations is further improved. As shown in Table 2, compared with EFT, the performance of both decision-making and question generation decline when RD candidate rule texts are removed, highlighting the effectiveness of RD candidate rule

Model	Seen				Unseen				Parameters
	Micro	Macro	$F1_{BLEU1}$	$F1_{BLEU4}$	Micro	Macro	$F1_{BLEU1}$	$F1_{BLEU4}$	
MUDERN	88.1	88.1	62.6	57.8	66.4	66.0	33.1	24.3	330M
OSCAR	–	–	64.6	59.6	–	–	**34.9**	**25.1**	330M
EFT	**92.4**	**92.3**	**83.4**	**81.3**	**68.4**	**68.2**	**34.9**	24.0	**220M**

Table 3: The comparison of question generation on the seen and unseen splits on test-set.

texts in enhancing the information-seeking abilities of our fused answer generation decoder. By removing RD candidate rule texts, the micro-acc is decreased by 0.5, the macro-acc is decrease by 0.4, the $F1_{BLEU1}$ is decreased by 1.7, and the $F1_{BLEU4}$ is decreased by 2.0. The above results emphasize the indispensability of RD candidate rule texts.

Analysis of Activated Entailment Reasoning EFT-wo/s+a trains this model additionally remove activated entailment reasoning. As illustrated in Table 2, compared with EFT-wo/s+a, the performance of both decision-making and question generation are dropped without activated entailment reasoning, the micro-acc is decreased by 2.2, the macro-acc is decrease by 2.3, the $F1_{BLEU1}$ is decreased by 1.4, and the $F1_{BLEU4}$ is decreased by 0.7. The above results suggest that the implicit reasoning of conversational machine reading comprehension could be enhanced by introducing explicit fine-grained supervised signal in a global understanding manner.

Analysis of Order Information Protection The order of fused representation used in fused answer generation decoder may lead to information leakage and affect the reasoning ability of the decoder. In order to avoid the problem of poor information seeking ability caused by excessive learning of position information of the model, we randomly shuffle the order of fused representation to protect the order information in the training stage. As illustrated in Table 2, compared with EFT-wo/s+a, EFT-wo/s+a+i decrease the performance of both decision-making and question generation without order information protection, the micro-acc is decreased by 0.5 , the macro-acc is decrease by 0.6, the $F1_{BLEU1}$ is decreased by 1.2, and the $F1_{BLEU4}$ is decreased by 1.3. The above results indicates the importance of order information protection.

Analysis of Fused Generation Fused Generation is utilized to introduce the ability to process multiple rule contextualized information. The multiple rule contextualized information are fused as a single fused information. EFT-wo/s+a+i+f trains this model without multi rule fused answer generation, only top1 retrieved rule text is considered. In this manner, the performance is limited with the retrieval performance. Compared with EFT-wo/s+a+i+f, the performance of both decision-making and question generation of EFT-wo/s+a+i are significantly improved by introducing fused answer generation strategy, the micro-acc is increased by 9.2 , the macro-acc is increased by 8.9, the $F1_{BLEU1}$ is increased by 12.0, and the $F1_{BLEU4}$ is increased by 11.2. The above results suggests the necessity of fused answer generation strategy.

Model	Yes		No		Inquire	
	dev	test	dev	test	dev	test
E^3	58.5	58.5	61.8	60.1	66.5	66.4
EMT	56.9	55.4	68.6	65.6	74.0	73.6
DISCERN	61.7	65.8	61.1	61.8	77.3	73.6
MP-RoBERTa	68.9	72.6	80.8	74.2	69.5	63.4
MUDERN	73.9	76.4	80.8	72.2	81.7	77.4
EFT	**80.1**	**81.2**	**83.2**	**75.6**	**88.2**	**78.7**

Table 4: Class-wise decision making accuracy among `Yes`, `No` and `Inquire` on the dev and test set of OR-ShARC.

5 Conclusion

In this paper, we propose a novel end-to-end framework, called EFT, to bridge the information gap between decision-making and question generation through the shared entailment representation in a global understanding manner. Extensive experimental results on the OR-ShARC benchmark demonstrate the effectiveness of our proposed framework EFT. In our analysis, the implicit reasoning ability of both decision-making and question generation is significantly improved by sharing external explicit entailment knowledge through our novel framework EFT.

Limitations

As shown in Table 3, the results demonstrate the effectiveness of our proposed EFT, but the performance of the unseen subset is still limited by comparing it with the performance of seen subset, which suggests plenty of room for improvement. Data augmentation or generalization methods based on semi-supervised methods could be effective to solve the problem in the future.

Acknowledgements

The work is supported by National Key R&D Plan (No. 2020AAA0106600), National Natural Science Foundation of China (No.62172039, U21B2009 and 62276110).

References

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*, Brussels, Belgium. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

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*, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Yifan Gao, Chien-Sheng Wu, Shafiq Joty, Caiming Xiong, Richard Socher, Irwin King, Michael Lyu, and Steven C.H. Hoi. 2020a. 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*.

Yifan Gao, Chien-Sheng Wu, Jingjing Li, Shafiq Joty, Steven C.H. Hoi, Caiming Xiong, Irwin King, and Michael Lyu. 2020b. 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.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

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)*, Hong Kong, China. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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. Association for Computational Linguistics.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. Segbot: a generic neural text segmentation model with pointer network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4166–4172.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. In *International Conference on Learning Representations*.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2021. Dialogue graph modeling for conversational machine reading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online. Association for Computational Linguistics.

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).

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.

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*.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

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)*, Online. Association for Computational Linguistics.

Xiao Zhang, Heyan Huang, Zewen Chi, and Xian-Ling Mao. 2022. ET5: A novel end-to-end framework for conversational machine reading comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea. International Committee on 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*.

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*.

Appendix

DPR-R	Top1	Top5	Top10	Top20
Dev	54.5	93.4	99.2	99.5
Seen Only	96.8	100.0	100.0	100.0
Unseen Only	19.5	87.9	98.5	99.0
Test	77.5	93.5	96.3	98,8
Seen Only	97.6	100.0	100.0	100.0
Unseen Only	62.8	88.8	93.7	97.9

Table 5: Retrieval Results of DPR-R.

A The Performance of Retriever (DPR-R)

Table 6 presents the detailed performance of DPR-R, including the performance of Top-k on dev set and test set. Different from previous methods (Zhang et al., 2021) that utilize DRP but only use TF-IDF retrieved negatives, we use random negatives sampled from seen rule texts in knowledge base. Experimental results illustrate that DPR-R outperforms DPR by 16.5% in top5 accuracy on test set, and reaches competitive results with TF-IDF+DPR.

B The Performance of Retriever (DPR-R) on Subset

We further analyzed the performance of DPR-R on the seen and unseen subset. As shown in Table 5, experimental results demonstrate the effectiveness of DPR-R on seen sets, the top1 accuracy reached 97.6 on the seen subset of test set. But the performance still have a large latent space of improvement on unseen sets.

C Additional Experiment Details

We implement EFT with the PyTorch[4] library and using pre-trained Transformers from the Hugging Face[5] repositories. The retriever DPR-R is based on the DPR[6] repositories. The data of OR-ShARC are from the OR-ShARC[7] repository. The above repositories provide the data, models and licenses. The whole training process takes about several hours on eight Nvidia A100 GPUs.

[4] github.com/pytorch/pytorch
[5] github.com/huggingface/transformers
[6] github.com/facebookresearch/DPR
[7] github.com/YifanGao/open_retrieval_conversational_machine_reading

Model	Dev Set				Test Set			
	Top1	Top5	Top10	Top20	Top1	Top5	Top10	Top20
TF-IDF	53.8	83.4	94.0	96.6	66.9	90.3	94.0	96.6
DPR	48.1	74.6	84.9	90.5	52.4	80.3	88.9	92.6
TF-IDF + DPR	**66.3**	90.0	92.4	94.5	**79.8**	**95.4**	**97.1**	97.5
DPR-R(ours)	54.5	**93.4**	**99.2**	**99.5**	77.5	93.5	96.3	**98.8**

Table 6: Comparison of the open-retrieval methods.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section after section 5

☑ A2. Did you discuss any potential risks of your work?
Limitations section after section 5

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section1

☑ A4. Have you used AI writing assistants when working on this paper?
We use chatGPT to verify the grammar of our abstract.

B ☑ Did you use or create scientific artifacts?

Appendix.C

☑ B1. Did you cite the creators of artifacts you used?
Appendix.C

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix.C

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix.C and Section 4.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Table 1, Table2 in Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix.C and Section 4.2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.