DISCERN: Discourse-Aware Entailment Reasoning Network for Conversational Machine Reading

Yifan Gao†, Chien-Sheng Wu‡, Jingjing Li†, Shafiq Joty†,§, Steven C.H. Hoi‡, Caiming Xiong‡, Irwin King†, and Michael R. Lyu†
† The Chinese University of Hong Kong
‡ Salesforce Research
§ Nanyang Technological University
†{yfgao,lijj,king,lyu}@cse.cuhk.edu.hk
‡{cswu,sjoty,shoi,cxiong}@salesforce.com

Abstract

Document interpretation and dialog understanding are the two major challenges for conversational machine reading. In this work, we propose DISCERN, a discourse-aware entailment reasoning network to strengthen the connection and enhance the understanding for both document and dialog. Specifically, we split the document into clause-like elementary discourse units (EDU) using a pre-trained discourse segmentation model, and we train our model in a weakly-supervised manner to predict whether each EDU is entailed by the user feedback in a conversation. Based on the learned EDU and entailment representations, we either reply to the user our final decision “yes/no/irrelevant” of the initial question, or generate a follow-up question to inquiry more information. Our experiments on the ShARC benchmark (blind, held-out test set) show that DISCERN achieves state-of-the-art results of 78.3% macro-averaged accuracy on decision making and 64.0 BLEU1 on follow-up question generation. Code and models are released at https://github.com/Yifan-Gao/Discern.

1 Introduction

Conversational Machine Reading (CMR) is challenging because the rule text may not contain the literal answer, but provide a procedure to derive it through interactions (Saeidi et al., 2018). In this case, the machine needs to read the rule text, interpret the user scenario, clarify the unknown user’s background by asking questions, and derive the final answer. Taking Figure 1 as an example, to answer the user whether he is suitable for the loan program, the machine needs to interpret the rule text to know what are the requirements, understand he meets “American small business” from the user scenario, ask follow-up clarification questions about “for-profit business” and “not get financing from other resources”, and finally it concludes the answer “Yes” to the user’s initial question.

Existing approaches (Zhong and Zettlemoyer, 2019; Sharma et al., 2019; Gao et al., 2020) decompose this problem into two sub-tasks. Given the rule text, user question, user scenario, and dialog history (if any), the first sub-task is to make a decision among “Yes”, “No”, “Inquire” and “Irrelevant”. The “Yes/No” directly answers the user question and “Irrelevant” means the user question is unanswerable by the rule text. If the user-provided information (user scenario, previous dialogs) are not enough to determine his fulfillment or eligibility, an “Inquire” decision is made and the second sub-task is activated. The second sub-task is to capture the underspecified condition from the rule text and generate a follow-up question to clarify it.

Zhong and Zettlemoyer (2019) adopt BERT (De-
vlin et al., 2019) to reason out the decision, and propose an entailment-driven extracting and editing framework to extract a span from the rule text and edit it into the follow-up question. The current state-of-the-art model EMT (Gao et al., 2020) uses a Recurrent Entity Network (Henaff et al., 2017) with explicit memory to track the fulfillment of rules at each dialog turn for decision making and question generation.

In this problem, document interpretation requires identification of conditions and determination of logical structures because rules can appear in the format of bullet points, in-line conditions, conjunctions, disjunctions, etc. Hence, correctly interpreting rules is the first step towards decision making. Another challenge is dialog understanding. The model needs to evaluate the user’s fulfillment over the conditions, and jointly consider the fulfillment states and the logical structure of rules for decision making. For example, disjunctions and conjunctions of conditions have completely different requirements over the user’s fulfillment states. However, existing methods have not considered condition-level understanding and reasoning.

In this work, we propose **DISCERN**: Discourse-Aware Entailment Reasoning Network. To better understand the logical structure of a rule text and to extract conditions from it, we first segment the rule text into clause-like elementary discourse units (EDUs) using a pre-trained discourse segmentation model (Li et al., 2018). Each EDU is treated as a condition of the rule text, and our model estimates its entailment confidence scores over three states: **ENTAILMENT**, **CONTRADICTION** or **NEUTRAL** by reading the user scenario description and existing dialog. Then we map the scores to an entailment vector for each segmented condition. Then it reasons out the decision by considering the logical structure of the rule text and the fulfillment of each condition.

**2 DISCERN Model**

DISCERN answers the user question through a three-step process shown in Figure 2:

1. First, **DISCERN** segments the rule text into individual conditions using discourse segmentation.
2. Taking the user-provided information including the user question, user scenario and dialog history as inputs, **DISCERN** predicts the entailment state and maps it to an entailment vector for each segmented condition. Then it reasons out the decision by considering the logical structure of the rule text and the fulfilment of each condition.
3. Finally, if the decision is “Inquire”, **DISCERN** generates a follow-up question to clarify the underspecified condition in the rule text.

**2.1 Rule Segmentation**

The goal of rule segmentation is to understand the logical structure of the rule text and parse it into individual conditions for the ease of entailment reasoning. Ideally, each segmented unit should contain at most one condition. Otherwise, it will be ambiguous to determine the entailment state for that unit. Determining conditions is easy when they appear as bullet points, but in most cases (65% samples in the ShARC dataset), one rule sentence may contain several in-line conditions as exemplified in Figure 2. To extract these in-line conditions, we find discourse segmentation in discourse parsing to be useful. In the Rhetorical Structure Theory or RST (Mann and Thompson, 1988) of discourse parsing, texts are first split into a sequence of clause-like units called elementary discourse units (EDUs). We utilize an off-the-shelf discourse segmenter (Li et al., 2018) to break the rule text into a sequence of clause-like units called elementary discourse units (EDUs). We utilize an off-the-shelf discourse segmenter (Li et al., 2018) to break the rule text into a sequence of EDUs. The segmenter uses a pointer network and achieves 92.2% F-score with Glove vectors and 95.55% F-score with ELMo embeddings on the standard RST benchmark testset.
Step ①  Rule Text: If a worker has taken more leave than they’re entitled to, their employer must not take money from their final pay unless it’s been agreed beforehand in writing.

[If a worker has taken more leave than they’re entitled to, their employer must not take money from their final pay unless it’s been agreed beforehand in writing.]

Discourse Segmentation

Step ②  Rule Text: If a worker has taken more leave than they’re entitled to, their employer must not take money from their final pay unless it’s been agreed beforehand in writing.

| Rule Text | User Question | User Scenario | Dialog History |
|-----------|---------------|---------------|----------------|
| [CLS] EDU1 [CLS] EDU2 [CLS] EDU3 [SEP] | [CLS] Ques. [SEP] | [CLS] Scen. [SEP] | [CLS] QA1 [SEP] [CLS] QA2 [SEP] |
| e1 |
| e2 |
| e3 |
| uQ |
| uS |
| u1 |
| uM |

RoBERTa

Inter-sentence Transformer Layers

V_E (Entailment)  V_C (Contradiction)  V_N (Neutral)

Mapping

Decision Classifier

Yes  No  Irrelevant  Inquire

If ‘Inquire’, ask a follow-up question

Step ③  Follow-up Question

Figure 2: The overall diagram of our proposed DISCERN. DISCERN first segments the rule text into several elementary discourse units (EDUs) as conditions (Section 2.1). Then, taking the segmented conditions, user question, user scenario, and dialog history as inputs, DISCERN reasons out the decision among “Yes”, “No”, “Irrelevant” and “Inquire” (Section 2.2). If the decision is “Inquire”, the question generation model asks a follow-up question (Section 2.3). (Best viewed in color)

which is close to human agreement of 98.3% F-score (Joty et al., 2015; Lin et al., 2019b). As exemplified in Figure 2 Step ①, the rule sentence is broken into three EDUs, in which two conditions (“If a worker has taken more leave than they’re entitled to”, “unless it’s been agreed beforehand in writing”) and the outcome (“their employer must not take money from their final pay”) are split out precisely. For rule texts which contain bullet points, we directly treat these bullet points as conditions.

2.2 Decision Making via Entailment Reasoning

Encoding. As shown in Figure 2 Step ②, inputs to DISCERN include the segmented conditions (EDUs) in the rule text, user question, user scenario, and follow-up question-answer pairs in dialog history, each of which is a sequence of tokens. In order to get the sentence-level representations for all individual sequences, we insert an external [CLS] symbol at the start of each sequence, and add a [SEP] symbol at the end of every type of inputs. Then, DISCERN concatenates all sequences together, and uses RoBERTa (Liu et al., 2019) to encode the concatenated sequence. The encoded [CLS] token represents the sequence that follows it. In this way, we extract sentence-level representations of conditions (EDUs) as e_1, e_2, ..., e_N, and also the representations of the user question u_Q, user scenario u_S, and M turns of dialog history u_1, ..., u_M. All these vectorized representations are of d dimensions (768 for RoBERTa-base).
Entailment Prediction. In order to reason out the correct decision for the user question, it is necessary to figure out the fulfillment of conditions in the rule text. We propose to formulate the fulfillment prediction of conditions into a multi-sentence entailment task. Given a sequence of conditions (premises) and a sequence of user-provided information (hypotheses), a system should output ENTAILMENT, CONTRADICTION or NEUTRAL for each condition listed in the rule text. In this context, NEUTRAL indicates that the condition has not been mentioned from the user information.

We utilize an inter-sentence transformer encoder (Vaswani et al., 2017) to predict the entailment states for all conditions simultaneously. Taking all sentence-level representations \( [e_1; e_2; \ldots; e_N; u_Q; u_S; u_1; \ldots; u_M] \) as inputs, the \( L \)-layer transformer encoder makes each condition attend to all the user-provided information to predict whether the condition is entailed or not. We also allow all conditions can attend to each other to understand the logical structure of the rule text.

Let the transformer encoder output of the \( i \)-th condition as \( \tilde{e}_i \), we use a linear transformation to predict its entailment state:

\[
c_i = W_c \tilde{e}_i + b_c \in \mathbb{R}^3,
\]

where \( c_i \in [c_E,i, c_C,i, c_N,i] \in \mathbb{R}^3 \) contains confidence scores of three entailment states ENTAILMENT, CONTRADICTION, NEUTRAL for the \( i \)-th condition in the rule text.

Since there are no ground truth entailment labels for individual conditions, we adopt a heuristic approach similar to Gao et al. (2020) to get the noisy supervision signals. Given the rule text, we first collect all associated follow-up questions in the dataset. Each follow-up question is matched to a segmented condition (EDU) in the rule text which has the minimum edit distance. For conditions in the rule text which are mentioned by follow-up questions in the dialogue history, we label the entailment state of a condition as Entailment if the answer for its mentioned follow-up question is Yes, and label the state of this condition as Contradiction if the answer is No. The remaining conditions not covered by any follow-up question are labeled as Neutral. Let \( r \) indicate the correct entailment state. The entailment prediction is weakly supervised by the following cross entropy loss, normalized by total number of \( K \) conditions in a batch:

\[
\mathcal{L}_{\text{entail}} = -\frac{1}{K} \sum_{i=1}^{K} \log \text{softmax}(c_i)_r
\]

Decision Making. After knowing the entailment state for each condition in the rule text, the remaining challenge for decision making is to perform logical reasoning over different rule types such as disjunction, conjunction, and conjunction of disjunctions. To achieve this, we first design three \( d \)-dimension entailment vectors \( V_E \) (Entailment), \( V_C \) (Contradiction), \( V_N \) (Neutral), and map the predicted entailment confidence scores of each condition to its vectorized entailment representation:

\[
V_{\text{EDU},i} = \sum_{k \in \{E,C,N\}} c_{k,i} V_k \in \mathbb{R}^d,
\]

These entailment vectors are randomly initialized and then learned during training. Finally, DISCERN jointly considers the logical structure of rules \( \tilde{e}_i \) and the entailment representations \( V_{\text{EDU},i} \) of conditions to make a decision:

\[
\alpha_i = w_{\alpha}^T [V_{\text{EDU},i}; \tilde{e}_i] + b_{\alpha} \in \mathbb{R}^1
\]

\[
\tilde{\alpha}_i = \text{softmax}(\alpha_i) \in [0, 1]
\]

\[
g = \sum_i \tilde{\alpha}_i [V_{\text{EDU},i}; \tilde{e}_i] \in \mathbb{R}^{2d}
\]

\[
z = W_z g + b_z \in \mathbb{R}^4
\]

where \( [V_{\text{EDU},i}; \tilde{e}_i] \) denotes the vector concatenation, \( \alpha_i \) is the attention weight for the \( i \)-th condition that determines whether the \( i \)-th condition should be taken into consideration for the final decision. \( z \in \mathbb{R}^4 \) contains the predicted scores for all four possible decisions “Yes”, “No”, “Inquire” and “Irrelevant”. Let \( l \) indicate the correct decision, \( z \) is supervised by the following cross entropy loss:

\[
\mathcal{L}_{\text{dec}} = -\log \text{softmax}(z)_l
\]

The overall loss for the Step 2 decision making is the weighted-sum of decision loss and entailment prediction loss:

\[
\mathcal{L} = \mathcal{L}_{\text{dec}} + \lambda \mathcal{L}_{\text{entail}}
\]

2.3 Follow-up Question Generation

If the predicted decision is “Inquire”, the follow-up question generation model is activated, as shown in Step 3 of Figure 2. It extracts an underspecified span from the rule text which is uncovered from
the user’s feedback, and rephrases it into a well-formed question. Existing approaches put huge efforts in extracting the underspecified span, such as entailment-driven extracting and ranking (Zhong and Zettlemoyer, 2019) or coarse-to-fine reasoning (Gao et al., 2020). However, we find that such sophisticated modelings may not be necessary, and we propose a simple but effective approach here.

We split the rule text into sentences and concatenate the rule sentences and user-provided information into a sequence. Then we use RoBERTa to encode them into vectors grounded to tokens, as here we want to predict the position of a span within the rule text. Let \([t_{1,1}, \ldots, t_{1,s_1}; t_{2,1}, \ldots, t_{2,s_2}; \ldots; t_{N,1}, \ldots, t_{N,s_N}]\) be the encoded vectors for tokens from \(N\) rule sentences, we follow the BERTQA approach (Devlin et al., 2019) to learn a start vector \(w_s \in \mathbb{R}^d\) and an end vector \(w_e \in \mathbb{R}^d\) to locate the start and end positions, under the restriction that the start and end positions must belong to the same rule sentence:

\[
\text{Span} = \arg \max_{i,j,k} (w_s^\top t_{k,i} + w_e^\top t_{k,j})
\]

where \(i, j\) denote the start and end positions of the selected span, and \(k\) is the sentence which the span belongs to. The training objective is the sum of the log-likelihoods of the correct start and end positions. To supervise the span extraction process, the noisy supervision of spans are generated by selecting the span which has the minimum edit distance with the to-be-asked question. Lastly, following Gao et al. (2020), we concatenate the rule text and span as the input sequence, and finetune UniLM (Dong et al., 2019), a pre-trained language model to rephrase it into a question.

3 Experiments

3.1 Experimental Setup

Dataset. ShARC (Saeidi et al., 2018) dataset is the current benchmark to test entailment reasoning in conversational machine reading\(^1\). The dataset contains 948 rule texts clawed from 10 government websites, in which 65\% of them are plain text with in-line conditions while the rest 35\% contain bullet-point conditions. Each rule text is associated with a dialog tree (follow-up QAs) that considers all possible fulfillment combinations of conditions. In the data annotation stage, parts of the dialogs are paraphrased into the user scenario. These parts of dialogs are marked as evidence which should be extracted (entailed) from the user scenario, and are not provided as inputs for evaluation. The inputs to the system are the rule text, user question, user scenario, and dialog history (if any). The output is the answer among Yes, No, Irrelevant, or a follow-up question. The train, development, and test dataset sizes are 21890, 2270, and 8276, respectively.

Evaluation Metrics. The decision making sub-task uses macro- and micro- accuracy of four classes “Yes”, “No”, “Irrelevant”, “Inquire” as metrics. For the question generation sub-task, we evaluate models under both the official end-to-end setting (Saeidi et al., 2018) and the recently proposed oracle setting (Gao et al., 2020). In the official setting, the BLEU score (Papineni et al., 2002) is calculated only when both the ground truth decision and the predicted decision are “Inquire”, which makes the score dependent on the model’s “Inquire” predictions. For the oracle question generation setting, models are asked to generate a question when the ground truth decision is “Inquire”.

Implementation Details. For the decision making sub-task, we finetune RoBERTa-base model (Wolf et al., 2019) with Adam (Kingma and Ba, 2015) optimizer for 5 epochs with a learning rate of 5e-5, a warm-up rate of 0.1, a batch size of 16, and a dropout rate of 0.35. The number of inter-sentence transformer layers \(L\) and the loss weight \(\lambda\) for entailment prediction are hyperparameters. We try 1, 2, 3 for \(L\) and 1.0, 2.0, 3.0, 4.0, 5.0 for \(\lambda\), and find the best combination is \(L = 2, \lambda = 3.0\), based on the development set results. For the question generation sub-task, we train a RoBERTa-base model to extract spans under the same training scheme above, and finetune UniLM (Dong et al., 2019) 20 epochs for question rephrasing with the same training scheme above, and finetune UniLM (Dong et al., 2019) 20 epochs for question rephrasing with a batch size of 16, a learning rate of 2e-5, and a beam size 10 for decoding in the inference stage. We repeat 5 times with different random seeds for all experiments on the development set and report the average results along with their standard deviations. It takes two hours for training on a 4-core server with an Nvidia GeForce GTX Titan X GPU.

3.2 Results

Decision Making Sub-task. The decision making results in macro- and micro- accuracy on the blind, held out test set of ShARC are shown in Table 1. DISCERN outperforms the previous best

\(^1\)Leaderboard: https://sharc-data.github.io/leaderboard.html
Table 1: Performance on the blind, held-out test set of ShARC end-to-end task.

| Models                        | Micro Acc. | Macro Acc. | BLEU1 | BLEU4 |
|-------------------------------|------------|------------|-------|-------|
| Seq2Seq (Saeidi et al., 2018) | 44.8       | 42.8       | 34.0  | 7.8   |
| Pipeline (Saeidi et al., 2018)| 61.9       | 68.9       | 54.4  | 34.4  |
| BERTQA (Zhong and Zettlemoyer, 2019) | 63.6 | 70.8       | 46.2  | 36.3  |
| UrcaNet (Sharma et al., 2019) | 65.1       | 71.2       | 60.5  | 46.1  |
| BiSon (Lawrence et al., 2019) | 66.9       | 71.6       | 58.8  | 44.3  |
| E$^3$ (Zhong and Zettlemoyer, 2019) | 67.6 | 73.3       | 54.1  | 38.7  |
| EMT (Gao et al., 2020)        | 69.4       | 74.8       | 60.9  | 46.0  |
| EMT+entailment (Gao et al., 2020) | 69.1 | 74.6       | 63.9  | 49.5  |
| DISCERN (our single model)    | 73.2       | 78.3       | 64.0  | 49.1  |

Table 2: Class-wise decision prediction accuracy among “Yes”, “No”, “Inquire” and “Irrelevant” on the development set of ShARC.

| Models          | Yes | No  | Inq. | Irr. |
|-----------------|-----|-----|------|------|
| BERTQA          | 61.2| 61.0| 62.6 | 96.4 |
| E$^3$           | 65.9| 70.6| 60.5 | 96.4 |
| UrcaNet         | 63.3| 68.4| 58.9 | 95.7 |
| EMT             | 70.5| 73.2| 70.8 | 98.6 |
| DISCERN         | 71.9| 75.8| 73.3 | 99.3 |

Table 3: Oracle question generation performance on the development set of ShARC.

| Models          | BLEU1 ± | BLEU4 ± |
|-----------------|---------|---------|
| E$^3$           | 52.79±2.87 | 37.31±2.35 |
| E$^3$+UniLM     | 57.09±1.70 | 41.05±1.80 |
| EMT             | 62.32±1.62 | 47.89±1.58 |
| DISCERN (BERT)  | 64.13±0.43 | 50.73±0.72 |
| DISCERN         | 64.23±0.84 | 50.85±0.89 |

The decision making sub-task and the follow-up question generation sub-task do not share too many commonalities so the results are not improved for each task in their multi-task training. On the other hand, our question generation model is easy to optimize because this model is separately trained from the decision making one, which means there is no need to balance the performance between these two sub-tasks. Besides, RoBERTa backbone performs comparably with its BERT counterpart.

In our detailed analyses, we find DISCERN can locate the next questionable sentence with 77.2% accuracy, which means DISCERN utilizes the user scenario and dialog history well to locate the next underspecified condition. We try to add entailment prediction supervision to help DISCERN to locate the unfulfilled condition but it does not help. We also try to simplify our approach by directly fine-tuning UniLM to learn the mapping between concatenated input sequences and the follow-up clarification questions. However, the poor result (around 40 for BLEU1) suggests this direction still remains further investigations.

3.3 Ablation Study

Table 4 shows an ablation study of DISCERN for the decision making sub-task on the development set of ShARC, and we have the following observations:
\[\text{Table 4: Ablation Study of DISCERN for decision making on the development set of ShARC.}\]

| Models | Micro Acc. | Macro Acc. |
|--------|------------|------------|
| DISCERN | 74.97 ± 0.27 | 79.55 ± 0.35 |
| DISCERN (BERT) | 73.07 ± 0.21 | 77.77 ± 0.24 |
| DISCERN (w/o EDU) | 73.34 ± 0.22 | 78.25 ± 0.57 |
| DISCERN (w/o Trans) | 74.25 ± 0.36 | 78.78 ± 0.57 |
| DISCERN (w/o \(\tilde{e}\)) | 73.55 ± 0.26 | 78.19 ± 0.30 |
| DISCERN (w/o \(\epsilon_{\text{EDU}}\)) | 72.95 ± 0.23 | 77.53 ± 0.19 |

Presumably, the condition representations account for the logical forms of rule texts and entailment vectors contain the fulfillment states for these conditions.

\[\text{Table 5: Decision prediction accuracy categorized by logical types of rules on the ShARC development set.}\]

| Logical Type | # samples | Micro Acc. | Macro Acc. |
|--------------|-----------|------------|------------|
| Simple       | 569       | 82.78 ± 1.48 | 86.91 ± 1.31 |
| Disjunction  | 726       | 69.97 ± 1.85 | 75.89 ± 1.38 |
| Conjunction  | 698       | 74.47 ± 2.41 | 79.78 ± 1.74 |
| Other        | 277       | 73.29 ± 2.53 | 77.17 ± 1.54 |

3.4 Analysis of Logical Structure of Rules

To see how DISCERN understands the logical structure of rules, we evaluate the decision making accuracy according to the logical types of rule texts. Here we define four logical types: “Simple”, “Conjunction”, “Disjunction”, “Other”, which are inferred from the associated dialog trees. “Simple” means there is only one requirement in the rule text while “Other” denotes the rule text have complex logical structures, for example, a conjunction of disjunctions or a disjunction of conjunctions. Table 5 shows decision prediction results categorized by different logical structures of rules. DISCERN achieves the best performance on the “Simple” logical type which only needs to determine the single condition is satisfied or not. On the other hand, DISCERN does not perform well on rules in the format of disjunctions. We conduct further analysis on this category and find that the error comes from user scenario interpretation: the user has already provided his fulfillment in the user scenario but DISCERN fails to extract it. Detailed analyses are further conducted in the following section.

3.5 How Far Has the Problem Been Solved?

In order to figure out the limitations of DISCERN, and the current challenges of ShARC CMR, we disentangle the challenges of scenario interpretation and dialog understanding in ShARC by selecting different subsets, and evaluate decision making and entailment prediction accuracy on them.

Baseline. Because the classification for unanswerable questions (“irrelevant” class) is nearly solved (99.3% in Table 2), we create the baseline subset by removing all unanswerable examples from the development set. Results for this baseline are shown in ShARC (Answerable) of Table 6.
### Dialog History Subset
We first want to see how DISCERN understands dialog histories (follow-up QAs) without the influence of user scenarios. Hence, we create a subset of ShARC (Answerable) in which all samples have an empty user scenario. The performance over 224 such samples is shown in “Dialog History Subset” of Table 6. Surprisingly, the results on this portion of samples are much better than the overall results, especially for the entailment prediction (92.41% micro-accuracy).

### Scenario Subset
With the curiosity to see what is the bottleneck of our model, we test the model ability on scenario interpretation. Similarly, we create a “Scenario Subset” from ShARC (Answerable) in which all samples have an empty dialog history. Results in Table 6 (“Scenario Subset”) show that interpreting scenarios to extract the entailment information within is exactly the current bottleneck of DISCERN. We analyze 100 error cases on this subset and find that various types of reasoning are required for scenario interpretation, including numerical reasoning (15%), temporal reasoning (12%), and implication over common sense and external knowledge (46%). Besides, DISCERN still fails to extract user’s fulfillment when the scenarios paraphrase the rule texts (27%). Examples for each type of error are shown in Figure 3. Among three classes of entailment states, we find that DISCERN fails to predict ENTAILMENT or CONTRADICTION precisely – it predicts NEUTRAL in most cases for scenario interpretation, resulting in high micro-accuracy in entailment prediction but the macro-accuracy is poor. The decision accuracy is subsequently hurt by the entailment results.

### ShARC (Evidence)
Based on the above observation, we replace the user scenario in the ShARC (Answerable) by its evidence and re-evaluate the overall performance on these answerable questions. As described in Section 3.1 Dataset, the evidence is the part of dialogs that should be entailed from the user scenario. Table 6 shows that the model improves 11.38% in decision making micro-accuracy if no scenario interpretation is required, which validates our above observation.

## 4 Related Work

### Entailment Reasoning in Reading Comprehension
Understanding entailments (or implications) of text is essential in dialog and question answering systems. ROPES (Lin et al., 2019a) requires reading descriptions of causes and effects and applying them to situated questions, while ShARC (Saeidi et al., 2018), the focus of DISCERN, requires to understand rules and apply them to questions asked by users in a conversational manner. Most existing methods simply use BERT to classify the answer without considering the structures of rule texts (Zhong and Zettlemoyer, 2019; Sharma et al., 2019; Lawrence et al., 2019). Gao et al. (2020) propose Explicit Memory Tracker (EMT), which firstly addresses entailment-oriented reasoning. At each dialog turn, EMT recurrently tracks whether conditions listed in the rule texts have already been satisfied to make a decision.

In this paper, we also explicitly model entailment reasoning for decision making, but there are three key differences between our DISCERN and EMT: (1) we apply discourse segmentation to parse the rule text, which is extremely helpful because there are many in-line conditions in rules; (2) Our stacked inter-sentence transformer layers extract better features for entailment prediction, which could be seen as a generalization of their recurrent explicit memory tracker. (3) Different from their utilization of entailment prediction which is treated as multi-task learning for decision making, we directly build the dependency between entailment prediction states and the predicted decisions.

### Discourse Applications
Discourse analysis uncovers text-level linguistic structures (e.g., topic, coherence, co-reference), which can be useful for many downstream applications, such as coherent text generation (Bosselut et al., 2018) and text sum-

### Table 6
Decision making and entailment prediction results over different subsets of the ShARC development set.

| Dataset                  | Decision Making Micro Acc. | Decision Making Macro Acc. | Entailment Prediction Micro Acc. | Entailment Prediction Macro Acc. |
|--------------------------|-----------------------------|-----------------------------|---------------------------------|---------------------------------|
| ShARC (Answerable)       | 73.55±0.33                  | 74.46±0.27                  | 86.41±0.39                      | 81.13±0.39                      |
| Dialog History Subset    | 79.29±1.62                  | 76.37±1.95                  | 92.41±0.38                      | 90.12±0.68                      |
| Scenario Subset          | 63.50±1.58                  | 60.18±1.72                  | 82.76±0.46                      | 59.40±1.04                      |
| ShARC (Evidence)         | 84.93±0.29                  | 84.37±0.24                  | 91.46±0.68                      | 89.90±1.40                      |
| Error Type               | % | Example |
|-------------------------|---|---------|
| Numerical Reasoning     | 15| **Relevant Rule**: Each attachment must be less than 10MB.  
**Scenario**: The attachment right now isn’t less than 10MB, but I think I can compress so it becomes less than 10MB.  
**Question**: Can I upload the attachment?  
**Entailment State** Gold: Entailment; Predict: Contradiction |
| Temporal Reasoning      | 12| **Relevant Rule**: The Additional State Pension is an extra amount of money you could get on top of your basic State Pension if you’re: . . . * a woman born before 6 April 1953  
**Scenario**: I live with my husband. We both have worked all our lives. Both of us were born in 1950.  
**Question**: Can I get Additional State Pension?  
**Entailment State** Gold: Entailment; Predict: Contradiction |
| Commonsense Reasoning   | 46| **Relevant Rule**: Homeowners may apply for up to $200,000 to repair or replace their primary residence to its pre-disaster condition.  
**Scenario**: My home was flooded  
**Question**: Is this loan suitable for me?  
**Entailment State** Gold: Entailment; Predict: Unknown |
| Paraphrase Reasoning    | 27| **Relevant Rule**: The Montgomery GI Bill (MGIB) is an educational assistance program enacted by Congress to attract high quality men and women into the Armed Forces.  
**Scenario**: I applied and found out I can get a loan. My dad wants me to join the army, but I don’t. I’d rather go to school.  
**Question**: Does this program meet my needs?  
**Entailment State** Gold: Contradiction; Predict: Unknown |

Figure 3: Types of scenario interpretation errors in the development data based on 100 samples.

marization (Joty et al., 2019; Cohan et al., 2018; Xu et al., 2020). Recently, discourse information has also been introduced in neural reading comprehension. Mihaylov and Frank (2019) design a discourse-aware semantic self-attention mechanism to supervise different heads of the transformer by discourse relations and coreferring mentions. Different from their use of discourse information, we use it as a parser to segment surface-level in-line conditions for entailment reasoning.

5 Conclusion

In this paper, we present DISCERN, a system that does discourse-aware entailment reasoning for conversational machine reading. DISCERN explicitly builds the connection between entailment states of conditions and the final decisions. Results on the ShARC benchmark shows that DISCERN outperforms existing methods by a large margin. We also conduct comprehensive analyses to unveil the limitations of DISCERN and challenges for ShARC. In future, we plan to explore how to incorporate discourse parsing into the current decision making model for end-to-end learning. One possibility would be to frame them as multi-task learning with a common (shared) encoder. Another direction is leveraging current methods in question generation (Gao et al., 2019; Li et al., 2019) to improve the follow-up question generation sub-task since DISCERN is on par with the previous best model EMT.

Acknowledgments

We thank Max Bartolo and Patrick Lewis for evaluating our submitted models on the hidden test set. The work described in this paper was partially supported by following projects from the Research Grants Council of the Hong Kong Special Administrative Region, China: CUHK 2300174 (Collaborative Research Fund, No. C5026-18GF); CUHK 14210717 (RGC General Research Fund).

References

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 615–621, New Orleans, Louisiana. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Jing Li, Aixin Sun, and Shafiq R. Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4166–4172. ijcai.org.

Jingjing Li, Yifan Gao, Lidong Bing, Irwin King, and Michael R. Lyu. 2019. Improving question generation with the to the point context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3216–3226, Hong Kong, China. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Steven C.H. Hoi. 2020. 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.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Todor Mihaylov and Anette Frank. 2019. Discourse-aware semantic self-attention for narrative reading comprehension. 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 2541–2552, Hong Kong, China. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, TiiT 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 12–17, Florence, Italy. Association for Computational Linguistics.
Abhishek Sharma, Danish Contractor, Harshit Kumar, and Sachindra Joshi. 2019. Neural conversational qa: Learning to reason v.s. exploiting patterns. ArXiv, abs/1909.03759.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics - on what language model pre-training captures. ArXiv, abs/1912.13283.

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 30, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5021–5031, Online. 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.