Dialogue Graph Modeling for Conversational Machine Reading

Siru Ouyang[1,2,3,*], **Zhuosheng Zhang**[1,2,3,*], **Hai Zhao**[1,2,3,†]
[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
{oysr0926,zhangzs}@sjtu.edu.cn,zhaohai@cs.sjtu.edu.cn

Abstract

Conversational Machine Reading (CMR) aims at answering questions in complicated interactive scenarios. Machine needs to answer questions through interactions with users based on given rule document, user scenario and dialogue history, and even initiatively asks questions for clarification if necessary. Namely, the answer to the task needs a machine in the response of either *Yes*, *No*, *Irrelevant* or to raise a follow-up question for further clarification. To effectively capture multiple objects in such a challenging task, graph modeling is supposed to be adopted, though it is surprising that this does not happen until this work proposes a dialogue graph modeling framework by incorporating two complementary graph models, i.e., explicit discourse graph and implicit discourse graph, which respectively capture explicit and implicit interactions hidden in the rule documents. The proposed model is evaluated on the ShARC benchmark and achieves new state-of-the-art by first exceeding the milestone accuracy score of 80%. The source code of our paper is available at https://github.com/ozyyshr/DGM

1 Introduction

Training machines to understand documents is the major goal of machine reading comprehension (MRC) (Hermann et al., 2015; Hill et al., 2016; Rajpurkar et al., 2016; Nguyen et al., 2016; Joshi et al., 2017; Rajpurkar et al., 2018; Choi et al., 2018; Zhang et al., 2018; Reddy et al., 2019; Zhang et al., 2020c, 2021). Especially, in the recent challenging conversational machine reading (CMR) task, the machine is required to read and interpret the given rule document and the user scenario, ask clarification questions, and then make a final decision (Saeidi et al., 2018). As an example shown in Figure 1. The user posts the scenario and asks a question concerning whether the loan meets the needs. Since the user cannot know the rule document, the information he/she provided may not be sufficient for the machine to decide. Therefore, a series of follow-up questions are asked by the machine until it can finally make a conclusion.

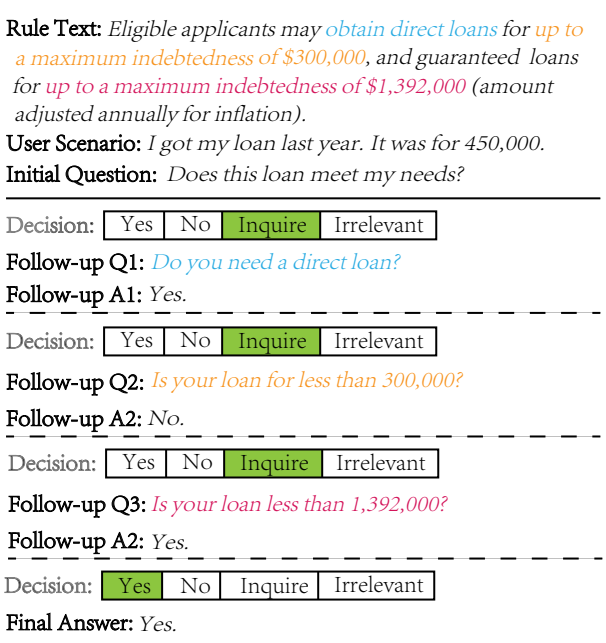


Figure 1: An example dialog from ShARC benchmark dataset (Saeidi et al., 2018). At each turn, the machine can give a decision regarding the initial question put up by the user. If the decision is *Inquire*, the machine will ask a clarification question to help with decision making. The corresponding rule document and the question are marked in the same color in the figure.

The major challenges for the conversational machine reading include the rule document interpretation, and reasoning with the background

* Equal contribution. †Corresponding author. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), Key Projects of National Natural Science Foundation of China (U1836222 and 61733011), Huawei-SJTU long term AI project, Cutting-edge Machine Reading Comprehension and Language Model. This work was supported by Huawei Noah's Ark Lab.

knowledge, e.g., the provided rule document, user scenario and the input question. Existing works (Zhong and Zettlemoyer, 2019; Lawrence et al., 2019; Verma et al., 2020; Gao et al., 2020a,b) have made progress in improving the reasoning ability by modeling the interactions among rule document, user scenario and the other elements implicitly. As for rule document interpretation, most existing approaches simply split the rule document into several rule conditions to be satisfied. In general, they first track the entailment state of each rule condition for decision making and then form a certain under-specified rule span into a follow-up question.

However, the aforementioned cascaded methods tend to model in a holistic way, i.e. interpreting the rule document with other elements quite plainly, which have the following drawbacks. First, very little attention is paid to the inner dependencies of rule conditions such as the discourse structure and discourse relations (Qin et al., 2016, 2017; Bai and Zhao, 2018). Second, existing methods do not dig deep enough into mining the interactions between the rule document and other elements, especially user scenarios.

As seen, the interactions of elements in CMR is far more complicated than that of traditional MRC tasks. Therefore, we proposed a dialogue graph modeling (DGM) framework consisting of two complementary graphs to fully capture the complicated interactions among all the elements. Firstly, an explicit discourse graph is constructed by making use of discourse relations of elementary discourse units (EDUs) generated from rule documents to tackle explicit element interactions. User scenario representation is injected as a special global vertex, to bridge the interactions and capture the inherent dependency between the rule document and the user scenario information. Secondly, an implicit discourse graph is designed for digging the latent salient interactions among rule documents by decoupling and fusing mechanism. The two dialogue graphs compose the encoder of our model and feed fusing representations to the decoder for making decisions.

As to our best knowledge we are the first to explicitly model the relationships among rules and user scenario with Graph Convolutional Networks (GCNs) (Schlichtkrull et al., 2018). Experimental results show that our proposed model outperforms the baseline models in terms of official evaluation metrics and achieves the new state-of-the-art results on ShARC, the benchmark dataset for CMR (Saeidi et al., 2018). In addition, our model enjoys strong interpretability by modeling the process in an intuitive way.

2 Related Work

Conversational Machine Reading. Compared with traditional triplet-based MRC tasks that aim to answer questions by reading given document (Hermann et al., 2015; Hill et al., 2016; Rajpurkar et al., 2016; Nguyen et al., 2016; Joshi et al., 2017; Rajpurkar et al., 2018; Zhang et al., 2020a,b), our concerned CMR task (Saeidi et al., 2018) is more challenging as it involves rule documents, scenarios, asking clarification question, and making a final decision. The major differences lie in two sides: 1) machines are required to formulate follow-up questions for clarification before confident enough to make the decision, 2) machines have to make a question-related conclusion by interpreting a set of complex decision rules, instead of simply extracting the answer from the text. Existing works (Zhong and Zettlemoyer, 2019; Lawrence et al., 2019; Verma et al., 2020; Gao et al., 2020a,b) have made progress in improving the reasoning ability by modeling the interactions between the rule document and other elements. As a widely-used manner, the existing models commonly extracted the rule documents into individual rule items, and track the rule fulfillment for the dialogue states. As indicated in Gao et al. (2020b), improving the rule document representation remains a key factor to the overall model performance, because the rule documents are formed with a series of implicit, separable, and possibly interrelated rule items that the conversation should satisfy before making decisions. However, previous work only considered segmenting the discourse, and neglected the inner discourse structure/relationships between the EDUs (Gao et al., 2020b). Compared to existing methods, our method makes the first attempt to explicitly capture elaborate interactions among all the document elements, user scenarios and dialogue history updates.

Graph Modeling in MRC. Inspired by the impressive performance of GCN (Kipf and Welling, 2017; Luo and Zhao, 2020), efforts towards better performance on MRC utilizing GCNs have sprung up, such as BAG (Cao et al., 2019), GraphRel (Fu et al., 2019) and social information reasoning (Li and Goldwasser, 2019). Unlike the previous works

who just apply the graph framework mechanically to turn the entire passage or document into a graph, the discourse graph we proposed is delicately designed to mine the relationships of multiple elements in CMR task and to facilitate information flow over the graph.

3 Model

As illustrated in Figure 2, our model mainly consists of three parts to generate the final answer.

1. Rule document is segmented into rule EDUs, which is then tagged discourse relationship by a pre-trained discourse parser.

2. In the encoding phase, taking segmented and preprocessed rule document and user scenario as input, we build two graphs over the segments (EDUs), in which the explicit discourse graph captures the interactions among rules and user scenarios with the support of tagged discourse relationship, while the implicit discourse graph mines latent salient interactions from the raw rule document.

3. For decoding, an interaction layer takes the combined representation generated by both explicit and implicit discourse graph of rule EDUs, initial question, user scenario and dialog history as inputs, and maps it into an entailment state of each rule EDU. With these rule fulfillment situation, we can make a decision among *Yes*, *No*, *Inquire* and *Irrelevant*. Once the decision is made to be *Inquire*, the model generates a follow-up question to clarify the under-specified rule span in the rule document.

The complete training procedure of DGM is shown in Algorithm 1.

Algorithm 1: DGM Algorithm

Input: word embeddings $E = \{e_1, .., e_n\}$, dimension of word embeddings d, token ids of rule document I, discourse relation D of rule document, number of rule EDUs n

Output: Final decision in Yes/No/Irrelevant or a follow-up question

for *i in epochs* **do**
- build explicit discourse graph $G(E, D)$
- $G_{n\times d} \leftarrow GCN(G)$
- build implicit discourse graph by calculating adjacent matrix M_l, M_c
- get rule EDU representation $C_{n\times d}$ by Eq.(6) and (7)
- combined representation for `[RULE]` $\tilde{r_i} \leftarrow$ self-Attn$(C + G)_i$
- entailment state $f_i \leftarrow$ LINEAR$(\tilde{r_i})$
- make the decision z by Eq.(10) based on $\tilde{r_i}$ and f_i
- **if** z *is Inquire* **then**
  - generate follow-up question
- **return** z or follow-up question

3.1 Preprocessing

EDU Segmentation. We first separate the rule document into several units each containing exactly one condition. Here we follow DISCERN (Gao et al., 2020b) adopting the discourse segmenter (Li et al., 2018) to break the rule document into EDUs.

Discourse Relation. Unlike EDU segmentation which only concerns with constituency-based logical structures, discourse relation allows relations between the non-adjacent EDUs. There are in total 16 discourse relations according to STAC (Asher et al., 2016), namely, *comment, clarification-question, elaboration, acknowledgement, continuation, explanation, conditional, question-answer, alternation, question-elaboration, result, background, narration, correction, parallel and contrast*. We adopt a pre-trained discourse parser (Shi and Huang, 2019)[1] to decide the dependencies between EDUs and the corresponding relation types with the structured representation of each EDU.

3.2 Encoding Block

Embedding. We select the pre-trained language model (PrLM) model ELECTRA (Clark et al., 2020) for encoding. As shown in the figure, the input of our model includes rule document which has already be parsed into EDUs with explicit discourse relation tagging, user initial question, user scenario and the dialog history. Instead of inserting a `[CLS]` token before each rule EDU to get a sentence-level representation, we use `[RULE]` which is proved to enhance performance (Lee et al., 2020). Note that we also insert `[SEP]` between every two adjacent utterances.

Explicit Discourse Graph. We first construct the explicit discourse graph as a Levi graph (Levi, 1942) which turns the labeled edges into additional vertices. Suppose $G = (V, E, R)$ is the graph constructed in the following way: if utterance U_1 is the continuation of utterance U_2,

[1] This discourse parser gives a state-of-the-art performance on STAC so far

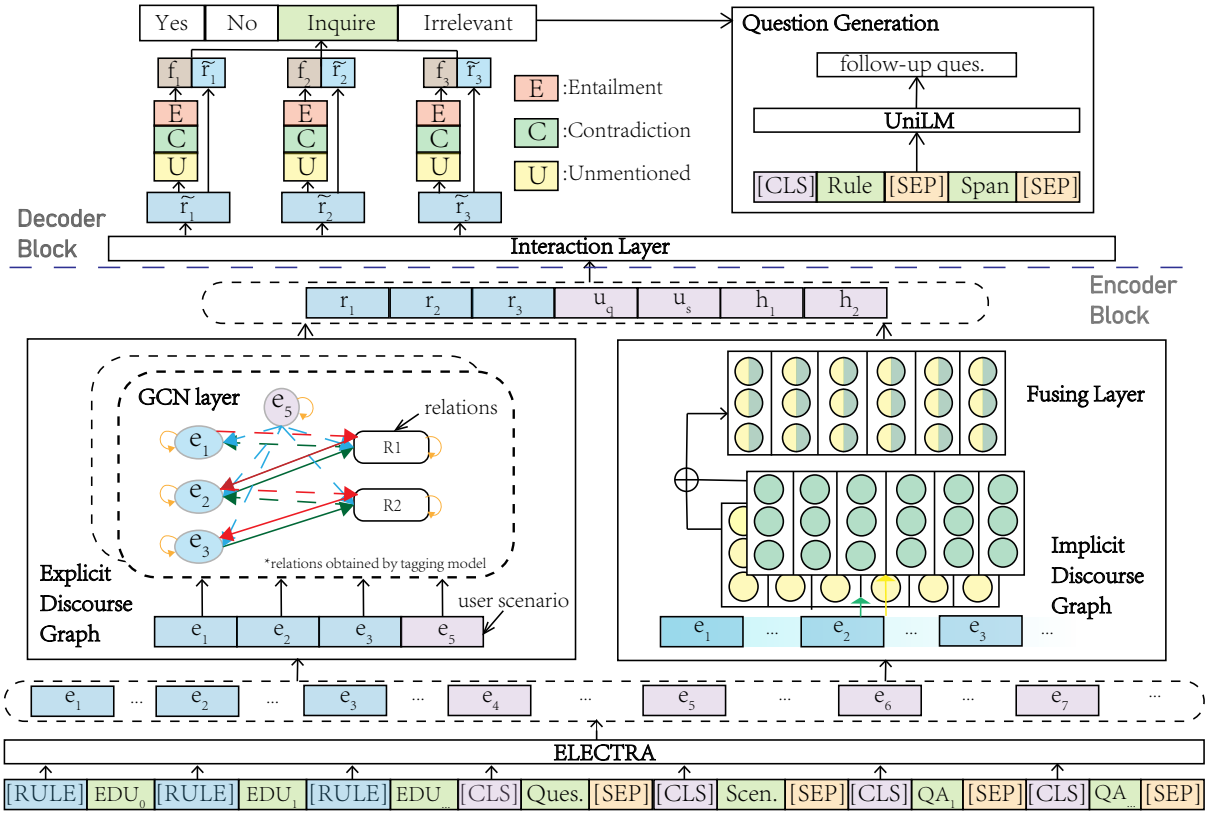


Figure 2: The overall structure for our proposed model. With segmented EDUs and tagged relations, the inputs including user initial question, user scenario and dialog history are sent for embedding and graph modeling to make the final decision. If the decision is *Inquire*, the question generation stage will be activated and use the under-specified span of rule document to generate a follow-up question.

we add a directed edge $e = (U_1, U_2)$ with relation R assigned to *Continuation*. The corresponding Levi graph can be expressed as $G = (V_L, E_L, R_L)$ where $V_L = V \cup R$. E_L is the set of edges with format $(U_1, Continuation)$ and $(Continuation, U_2)$. As for R_L, previous works such as (Marcheggiani and Titov, 2017; Beck et al., 2018) designed three types of edges $R_L = default, reverse, self$ to enhance information flow. Here with our settings, we extend it into six types: *default-in*, *default-out*, *reverse-in*, *reverse-out*, *self*, *global*, corresponding to the direction of the edges towards the relation vertices. An example of constructing Levi graph is shown in Figure 3. To construct the discourse structure of other elements, a global vertex representing user scenario is added and connected with all the other vertices.

We use a relational graph convolutional network (Schlichtkrull et al., 2018) to implement explicit discourse graph as the traditional GCN is not able to handle multi-relation graphs. For utterance and scenario vertices, we employ the encoding results of `[RULE]` and `[CLS]` in Section 3.1. For relation vertices, we look up in the embedding table to get the initial representation. Given the initial representation h_p^0 of every node v_p, the feed-forward or the message-passing process can be written as:

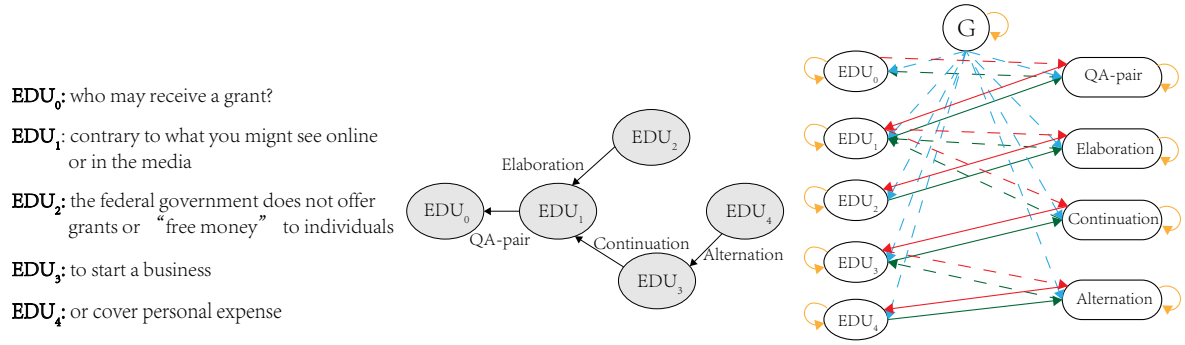
$$h_p^{(l+1)} = \text{ReLU}\Big(\sum_{r \in R_L} \sum_{v_q \in \mathcal{N}_r(v_p)} \frac{1}{c_{p,r}} w_r^{(l)} h_q^{(l)}\Big), \tag{1}$$

where $\mathcal{N}_r(v_p)$ denotes the neighbors of node v_p under relation r and $c_{p,r}$ is the number of those nodes. $w_r^{(l)}$ is the learnable parameters of layer l.

Because the total 16 relations cannot be treated equally, e.g. relation *Contrast* is much more important than the relation *Continuation*, we introduce the gating mechanism (Marcheggiani and Titov, 2017). The basic idea is to calculate a value between 0 and 1 for information passing control.

$$g_p^{(l)} = \text{Sigmoid}(h_p^{(l)} W_{r,g}), \tag{2}$$

where $W_{r,g}^{(l)}$ is a learnable parameter under relation type r of the l-th layer. Finally, the forward process



(a) separated EDUs (b) the original graph (c) the Levi graph

Figure 3: Processes turning a sample dialog into Levi graph representing discourse relations.

of gated GCN can be represented as:

$$h_p^{(l+1)} = \text{ReLU}(\sum_{r \in R_L} \sum_{v_q \in \mathcal{N}_r(v_p)} g_q^{(l)} \frac{1}{c_{p,r}} w_r^{(l)} h_q^{(l)}), \tag{3}$$

Implicit Discourse Graph. Implicit discourse graph aims at digging the salient latent interactions inside rule document. Each token i in rule EDU is represented as a vertex in the graph. We use adjacent matrices to express implicit discourse graph. Two types of matrices M_l and M_c are introduced standing for local and contextualized information:

$$M_l[i,j] = \begin{cases} 0 & \text{if } I_i = I_j, \\ -\infty & \text{otherwise.} \end{cases} \tag{4}$$

$$M_c[i,j] = \begin{cases} 0 & \text{if } I_i \neq I_j, \\ -\infty & \text{otherwise.} \end{cases} \tag{5}$$

where I_i is the index of token i in EDU. Thus the information containing in rule document are decoupled in two separate aspects. Using multi-head self-attention to encode the graph and denote the length of the whole rule document as s, embedding dimension as d, we will get the following:

$$G_i = \text{MHSA}(E, M_i),\ i \in \{l, c\}, \tag{6}$$

where $G_i \in \mathbb{R}^{s \times d}$ and E is the embedding result from PrLM. MHSA denotes the multi-head self-attention (Vaswani et al., 2017).

After enough interactions inside rule EDUs, we then fuse the information (Liu et al., 2020) of these two implicit discourse graphs-like items[2] above in a gated manner by considering both the original and graph encoding representation of rule document.

$$\begin{aligned} \tilde{E}_1 &= \text{ReLU}(\text{FC}([E, G_l, E - G_l, E \odot G_l])), \\ \tilde{E}_2 &= \text{ReLU}(\text{FC}([E, G_c, E - G_c, E \odot G_c])), \\ g &= \text{Sigmoid}(\text{FC}([\tilde{E}_1, \tilde{E}_2])), \\ C &= g \odot G_l + (1 - g) \odot G_c, \end{aligned} \tag{7}$$

where FC is the fully-connected layer and $C \in \mathbb{R}^{s \times d}$. We take the calculated result of the original `[RULE]` to stand for the updated rule EDUs from C, denoted as c_i.

3.3 Decoding Block

Interaction Layer. We use an interaction layer to attend to all available information so far to learn in a systematic way. A self-attention layer (Vaswani et al., 2017) is adopted here allowing all the rule EDUs and other elements to attend to each other. Let $[r_1, r_2, ...; u_q; u_s; h_1, h_2, ...]$ denote all the representations, r_i is the combined sentence-level representation of explicit and implicit discourse graph, u_q, u_s and h_i stand for the representation of user question, user scenario and dialog history respectively. After encoding, the output can be displayed as $[\tilde{r_1}, \tilde{r_2}, ...; \tilde{u_q}, \tilde{u_s}; \tilde{h_1}, \tilde{h}2...]$.

Decision Making. Similar to existing works (Zhong and Zettlemoyer, 2019; Gao et al., 2020a,b), we apply an entailment-driven approach for decision making. A linear transformation tracks the fulfillment state of each rule EDU among Entailment, Contradiction and Unmentioned. At last, the decision is made according to:

$$f_i = W_f \tilde{r_i} + b_f \in \mathbb{R}^3, \tag{8}$$

[2]Taking self-attention weights as edges connecting representations (as node), it can be seen as graph as well.

Model	Dev Set				Test Set			
	Decision Making		Question Gen.		Decision Making		Question Gen.	
	Micro	Macro	BLEU1	BLEU4	Micro	Macro	BLEU1	BLEU4
NMT (Saeidi et al., 2018)	-	-	-	-	44.8	42.8	34.0	7.8
CM (Saeidi et al., 2018)	-	-	-	-	61.9	68.9	54.4	34.4
BERTQA (Zhong and Zettlemoyer, 2019)	68.6	73.7	47.4	54.0	63.6	70.8	46.2	36.3
UcraNet (Verma et al., 2020)	-	-	-	-	65.1	71.2	60.5	46.1
BiSon (Lawrence et al., 2019)	66.0	70.8	46.6	54.1	66.9	71.6	58.8	44.3
E^3 (Zhong and Zettlemoyer, 2019)	68.0	73.4	67.1	53.7	67.7	73.3	54.1	38.7
EMT (Gao et al., 2020a)	73.2	78.3	67.5	53.2	69.1	74.6	63.9	**49.5**
DISCERN (Gao et al., 2020b)	74.9	79.8	65.7	52.4	73.2	78.3	**64.0**	49.1
DGM (ours)	**78.6**	**82.2**	**71.8**	**60.2**	**77.4**	**81.2**	63.3	48.4

Table 1: Results on the blind held-out test set and the dev set of ShARC end-to-end task. Micro and Macro stand for Micro Accuracy and Macro Accuracy respectively.

where f_i is the score predicted for the three labels of the i-th condition. This prediction is trained via a cross entropy loss for multi-classification problems:

$$\mathcal{L}_{entail} = -\frac{1}{N}\sum_{i=1}^{N} \log \text{softmax}(f_i)_r, \quad (9)$$

where r is the ground-truth state of fulfillment.

After obtaining the state of every rule, we are able to give a final decision towards whether it is *Yes*, *No*, *Inquire* or *Irrelevant* by attention.

$$\begin{aligned} \alpha_i &= w_\alpha^T[f_i; \tilde{r_i}] + b_\alpha \in \mathbb{R}^1, \\ \tilde{\alpha_i} &= \text{softmax}(\alpha)_i \in [0, 1], \\ z &= W_z \sum_i \tilde{\alpha_i}[f_i; \tilde{r_i}] + b_z \in \mathbb{R}^4, \end{aligned} \quad (10)$$

where α_i is the attention weight for the i-th decision and z has the score for all the four possible states. The corresponding training loss is:

$$\mathcal{L}_{decision} = -\log \text{softmax}(z)_l, \quad (11)$$

The overall loss for decision making is:

$$\mathcal{L} = \mathcal{L}_{decision} + \lambda \mathcal{L}_{entail}. \quad (12)$$

Qustion Generation. If the decision is made to be *Inquire*, the machine need to ask a follow-up question to further clarify. Question generation in this part is mainly based on the uncovered information in the rule document, and then that information will be rephrased into a question. We predict the position of an under-specified span within a rule document in a supervised way. Following Devlin et al. (2019), our model learns a start vector $w_s \in \mathbb{R}^d$ and end vector $w_e \in \mathbb{R}^d$ to indicate the start and end positions of the desired span:

$$span = \arg\min_{i,j,k}(w_s^T t_{k,i} + w_e^T t_{k,j}), \quad (13)$$

where $t_{k,i}$ denote the i-th token in the k-th rule sentence. The ground-truth span labels are generated by calculating the edit-distance between the rule span and the follow-up questions. Intuitively, the shortest rule span with the minimum edit-distance is selected to be the under-specified span. Finally, we concatenate the rule document and the predicted span as an input sequence to fine-tune UniLM (Dong et al., 2019) and generate the follow-up question.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct experiments on ShARC dataset, the current CMR benchmark[3] collected by Saeidi et al. (2018). It contains up to 948 dialog trees clawed from government websites. Those dialog trees are then flattened into 32,436 examples consisting of *utterance_id*, *tree_id*, *rule document*, *initial question*, *user scenario*, *dialog history*, *evidence* and the *decision*. It is worth noting that evidence is the information that we need to extract from user information and thus will not be given in the testing phase. The sizes of train, dev and test are 21,890, 2,270 and 8,276 respectively. We also showed the generalizability of our model on the Multi-Turn Dialogue Reasoning (MuTual) dataset (Cui et al., 2020), which has 8,678 multiple choice samples and is divided into 7,376, 651, 651 of train, dev and test sets respectively.

Evaluation. For the decision-making subtask, ShARC evaluates the Micro- and Macro- Acc. for the results of classification. If both the prediction and ground truth of decision is *Inquire*,

[3]Leaderboard can be found at website https://sharc-data.github.io/leaderboard.html

BLEU(Papineni et al., 2002) score (particularly BLEU1 and BLEU4) will be evaluated on the follow-up question generation subtask.

Implementation Details. For rule EDU relation prediction, we keep all the default parameters of the original discourse relation parser[4], with $F1$ score achieving 55. In the decision-making stage, we finetune an ELECTRA-based model. The dimension of hidden states is 1024 for both the encoder and decoder. The training process uses Adam (Kingma and Ba, 2015) for 5 epochs with learning rate set to 5e-5. We also use gradient clipping with a maximum gradient norm of 2, and a total batch size of 16. In the question generation stage, for the sake of consistency, we also use an ELECTRA-based model for span extraction. For UniLM, we finetune it with a batch size of 16, a learning rate of 2e-5 and beam size is set to 10 for inference. It takes 3-4 hours for training on a single TITAN RTX 2080Ti GPU (24GB memory).

4.2 Results

Table 1 shows the results of DGM and all the baseline models for the End-to-End task on the blind held-out test set of ShARC[5]. Evaluating results indicate that DGM outperforms the baselines in most of the metrics. In particular, DGM outperforms the previous state-of-the-art model DISCERN by 4.2% in Micro Acc. and 2.9% in Macro Acc.

To test the generality of DGM on other different PrLMs and to do a fair comparison with previous models, We alter the underlying PrLMs to other variants in DGM and the previous state-of-the-art model DISCERN respectively. The results on the dev set of ShARC are shown in Table 2. In the first place, DGM performs better than DISCERN on all the PrLMs, which indicates the all-round superiority of DGM. Additionally, results on ELECTRA is generally better than that of BERT and RoBERTa. This indicates that ELECTRA is an even better trained PrLM. By the aforementioned analysis, our DGM can generally perform well on widely-used PrLMs.

[4] https://github.com/shizhouxing/DialogueDiscourseParsing

[5] As indicated in (Gao et al., 2020a,b), the question generation results normally suffer from randomness. As the focus of this task is the decision making task like previous studies.

PrLMs	Micro Acc.		Macro Acc.	
	DISCERN	DGM	DISCERN	DGM
BERT_{base}	69.8	70.4	75.3	76.0
RoBERTa_{base}	74.9	75.8	79.8	80.2
ELECTRA_{base}	75.2	75.5	79.7	80.4
BERT_{large}	72.8	73.0	77.8	78.0
RoBERTa_{large}	76.1	76.6	80.6	81.0
ELECTRA_{large}	77.2	78.6	80.3	82.2

Table 2: Performance of DISCERN and DGM on different PrLMs on the dev set of ShARC.

In addition, Table 3 lists the class-wise classification accuracy of our model. Results demonstrate that our model performs quite satisfactorily for all classification subtasks, outperforming all other models in three of all four subtasks though a minor behind on the *Irrelevant* subtask. Compared to competent models, our model boosts the performance with a great gain to judge whether the user's requirements need further inquiry or are already fulfilled. It is worth noting that the *Inquire* subtask is the most fundamental one among all subtasks required by the concerned CMR. The superiority of our model for this core subtask shows that our DGM model indeed effectively captures the complicated interactions among all the concerned document rules and scenarios.

Models	Total	Yes	No	Inquire	Irrelevant
BERTQA	63.6	61.2	61.0	62.6	96.4
E^3	68.0	65.9	70.6	60.5	96.4
UrcaNet	65.9	63.3	68.4	58.9	95.7
EMT	73.2	70.5	73.2	70.8	98.6
DISCERN	75.2	71.9	75.8	73.3	**99.3**
DGM (ours)	**77.8**	**75.2**	**77.9**	**76.3**	97.8

Table 3: Class-wise decision prediction accuracy on the dev set of ShARC.

5 Analysis

5.1 Ablation Study

To investigate the impacts of different graphs, we conducted an ablation study on the decision-making subtask which is the vital part of our model, directly influencing the results afterward. Detailed results on the dev set of ShARC in Table 4 show that both the explicit and implicit discourse graph are indispensable as removing any one of them causes a performance drop (1-3 points) on both Macro Acc. and Micro Acc. Especially, these two metrics drop by a great margin as we remove the explicit discourse graph, which shows that explicit discourse relation reasoning is crucial in CMR.

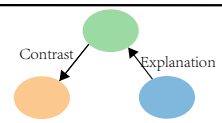
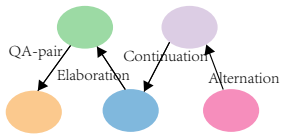
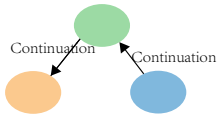
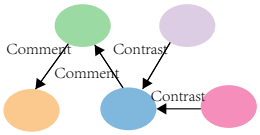
Example	Structure Type	Discourse Relation	Decision
Snippet: Export products made from endangered animals: (special rules)$_0$ If the animal is classied as B, C or D$_1$ you do not need to do anything$_2$ **Scenario:** I do not have medicare for this person. **Question:** Can I export products made from this animal?	Simple		Yes × No Irrelevant Inquire √
Snippet: Who may receive a grant?$_0$ Contrary to what you might see online in the media$_1$ the federal government does not offer grants or free money to individuals$_2$ to start a business$_3$ or cover personal expenses$_4$ **Scenario:** I live in Norway with my wife and children. **Question:** Can I get a grant from federal government?	Disjunction		Yes No × Irrelevant Inquire √
Snippet: In order to qualify for this benefit program$_0$ home-owners and renters must have sustained damage B, C or D$_1$ and be located in a disaster declared county$_2$ **Scenario:** (empty) **Question:** Do I qualify for this loan?	Conjunction		Yes No Irrelevant × Inquire √
Snippet: Going abroad$_0$ If your trip is going to last longer than 8 or 12 weeks, contact the Tax Credit Office within a month$_1$ your tex credits will end unless$_2$ you get UK benifits or State Pension$_3$ and you live in another European country with a child$_4$ **Scenario:** (empty) **Question:** Will my tex credit end?	Complex		Yes No √ Irrelevant Inquire ×

Figure 4: Examples selected from the dev set of ShARC where DISCERN fails but our model succeeds.

Also, we can see that adding the special token [RULE] indeed conduce to the performance.

Models	Macro Acc.	Micro Acc.
DGM	82.2	78.6
w/o Explicit Discourse Graph	79.8	75.2
w/o Implicit Discourse Graph	81.3	76.7
w/o both	77.3	71.7
w/o [RULE]	81.6	77.9

Table 4: Ablation study of our model for decision making subtask on the dev set of ShARC.

5.2 User Scenario Interpretation

In DGM, by injecting the user scenario as the global node in the explicit discourse graph we intend to improve the interpretation ability of our model with respect to user scenarios. To test the effect of the proposed model on scenario interpretation, we create a subset based on the dev set consisting of 761 samples that have user scenarios and an empty dialog history. The results on the decision-making subtask in Table 5 shows that our model can greatly improve the interpretation of user scenarios by surpassing DISCERN 11.8% and 14.8% of Macro Acc. and Micro Acc. respectively. In particular, DGM outperforms DISCERN by a large margin in every class of decision.

Models	Macro	Micro	Yes	No	Irrelevant	Inquire
DISCERN	63.5	60.2	35.3	50.7	100.0	67.8
DGM	**75.3**	**75.0**	**58.3**	**62.8**	**100.0**	**79.1**

Table 5: Results for decision making over user scenario subset of the ShARC dev set.

Manually analyzing the predicted results also indicates that DGM is capable of various reasoning including numerical reasoning, commonsense reasoning and rule document paraphrasing. For example, for numerical reasoning, given a scenario *"I plan on being away for five months before returning"*, DGM is able to match that with *"your tax credits will stop if you expect to be away for one year or more"* in the rule document.

5.3 Rule Document Interpretation

To see how DGM interpret the rule document, we analyzed the predictions from DGM and DISCERN on the dev set of ShARC to see how our model fixes the erroneous cases made by DISCERN.

We selected four types of rule structures and the representative examples are shown in Figure 4. It can be seen that the discourse relation tagged for the rule document can well represent the real relation in the discourse. For example, in the third case, *"homeowners and renters must have sustained damage B, C or D"* is the continuation of *"in order to qualify for this benefit program"* and *"be located in a disaster declared county"* is the continuation

of it. This kind of discourse relation informs that *"be located in a disaster declared county"* and *"have sustained damage B, C or D"* are two conditions one must obey to be qualified. Generally, *Continuation* may indicate that two rules have some conjunctive relations while *Alternation* denotes a disjunctive relation. All the relations together characterize the complex rule relations and thus are vital in decision making. Statistics regarding relations can be found in Table 6. The contextualized information containing in the rule document learned by the implicit discourse graph also contributes to the overall performance as it digs the semantically rich representations of rule EDUs.

Relation Types	Train Set	Dev Set
Comment	28756	2374
Clarification_question	330	69
Elaboration	639	82
Acknowledgement	6242	815
Continuation	7317	1090
Explanation	10831	1155
Conditional	1445	139
Question-answer_pair	1824	468
Alternation	896	323
Result	664	0
Correction	14	0
Contrast	16523	1595

Table 6: Statistics analysis of relation types of the train and dev on ShARC. "Comment", "Continuation", "Explanation" and "Contrast" constitutes the majority of the discourse relations.

5.4 Generalizability Evaluation

To verify DGM's generalizability and show that it can be smoothly applied to a broad type of QA tasks, We conducted experiments on a representative dialogue reasoning dataset MuTual. It is modified from Chinese high school English listening comprehension test data. It consists of 8860 annotated dialogues, namely, 7088 training samples, 886 developing samples and 886 testing samples. For each example, there is a dialogue history following by four candidate responses. Each candidate is relevant to the dialogue context but only one of them is logically correct. Our aim is to predict the correct answer given dialogue history and response candidates.

To apply DGM on MuTual, we first annotated the utterances of dialogue history of their discourse relations. Then pass the dialogue and the response candidates into a pre-trained language model to get the representation of each utterance. Armed with these representations and discourse relations, we are now able to construct the explicit discourse graph. Here, we set the global representation [CLS] (Devlin et al., 2019) of dialogue history as the global node. The implicit discourse graph can be constructed as Section 3.2 stated.

For the sake of computational efficiency, the maximum number of utterances is set to be 25. The concatenated context, response candidate in one sample is truncated or padded to be of length 256. We use ELECTRA as the PrLM and AdamW (Loshchilov and Hutter, 2019) as the optimizer for training. The batch size is 24 and the learning rate is 6e-6. We run a total of 3 epochs and select the model of the best results in the development set.

Table 7 displays the results on MuTual, which shows that DGM achieves a consistent improvement on the performance with respect to all the corresponding metrices.

Models	Dev Set			Test Set		
	R_4@1	R_4@2	MRR	R_4@1	R_4@2	MRR
ELECTRA	90.6	97.7	94.9	90.0	97.9	94.6
DGM	91.3	98.3	95.3	90.7	98.2	95.1

Table 7: Results on the dev and test set of MuTual dataset

6 Conclusions

In this paper, we presented a novel Dialogue Graph Modeling framework for Conversational Machine Reading. Our DGM consists of two complementary graphs which respectively capture both explicit and implicit interactions among multiple complicated elements in the challenging task, in which Explicit Discourse Graph is for extra knowledge learning with tagged EDU discourse relations while Implicit Discourse Graph helps with inside rule document understanding. Experiments on ShARC show the effectiveness by achieving a new state-of-the-art result. Our method may be smoothly applied to a broad type of QA tasks, such as our practice on the MuTual dataset that also achieves a consistent performance.

Acknowledgments

We would like to thank all the anonymous reviewers for their helpful comments and suggestions. Also thanks to Max Bartolo for evaluating our submitted models on the hidden test set.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583.

Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283.

Yu Cao, Meng Fang, and Dacheng Tao. 2019. BAG: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering. 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 357–362.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau 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 (EMNLP)*, pages 2174–2184.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

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.

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.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418.

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*, pages 935–945.

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.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

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*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks.

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.

Haejun Lee, Drew A. Hudson, Kangwook Lee, and Christopher D. Manning. 2020. SLM: Learning a discourse language representation with sentence unshuffling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1551–1562.

Friedrich Wilhelm Levi. 1942. *Finite geometrical systems: six public lectues delivered in February, 1940, at the University of Calcutta*. University of Calcutta.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604.

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.

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiaodong Zhou. 2020. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. *ArXiv*, abs/2009.06504.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ying Luo and Hai Zhao. 2020. Bipartite flat-graph network for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6408–6418.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1506–1515.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv:1611.09268v2*.

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.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2263–2270.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

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 (EMNLP)*, pages 2087–2097.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne vanden Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web*, pages 593–607.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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.

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.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020a. Semantics-aware BERT for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9628–9635. AAAI Press.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020b. Sg-net: Syntax-guided machine reading comprehension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9636–9643. AAAI Press.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *AAAI 2021*.

Zhuosheng Zhang, Hai Zhao, and Rui Wang. 2020c. Machine reading comprehension: The role of contextualized language models and beyond. *arXiv preprint arXiv:2005.06249*.

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.