Towards Coherent and Captivating Topic Transitions in Knowledge-Grounded Conversations

Anonymous ACL submission

Abstract

Knowledge-grounded conversations require skillful usage of knowledge to generate suitably diverse responses to keep user captivated while maintaining coherence to the dialogue context. However, current approaches that directly match knowledge with dialog context can result in capturing spurious correlations between knowledge and context, leading to either incoherent or mundane topic transitions in the generated dialogs that fail to engage. In this work, we introduce the Coherent and Captivating Topic Transition (C2T2) method to select the appropriate knowledge to be used in next response, resulting in topic transitions that are coherent to the ongoing conversations while providing adequate topic development for an engaging dialog. Our C2T2 employs transition-aware features designed to consider both historical contextual coherence as well as sequential topic development under a knowledge shifting constraint to select the next knowledge, thereby generating the response for an engaging conversation. We also designed a pointer network-based knowledge inference module to take into consideration of the relations among knowledge candidates during knowledge inference. Extensive experiments on two public benchmarks demonstrated the superiority of C2T2 on knowledge selection. Analysis on fine-grained knowledge selection accuracy also showed that C2T2 could better balance the topic adhesion and knowledge diversity in dialogs than existing approaches.

1 Introduction

A key challenge for open-domain dialog agents is to generate informative responses (Ghazvininejad et al., 2018) that leads to satisfy humans' need for information in communication. Knowledge-grounded conversations aim to leverage on external knowledge sources to generate informative response to engage the users, by learning from turn-level labelled knowledge (Dinan et al., 2019) resources that have become available recently. There

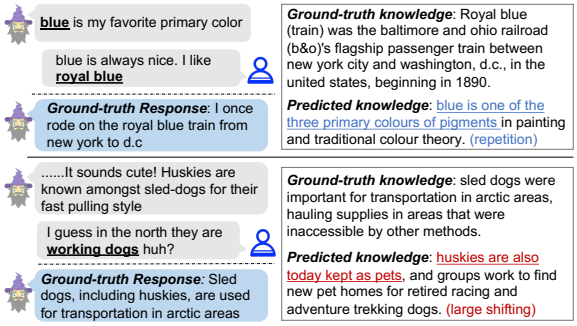

Figure 1: Two examples of inappropriate knowledge selection. Left part are dialogue histories and ground-truth responses with bold underlined words as topics. Right part compares ground-truth and predicted knowledge. The underlined phrases in the predicted knowledge illustrate either a repetition (top example) or a large shift (bottom example) of dialog topic transitions.

are two main steps in knowledge-grounded conversations: knowledge selection and response generation. The former selects a suitable knowledge from a knowledge pool that is appropriate for the next response, while the latter then generates a natural language response based the selected knowledge.

Knowledge selection is also known as the topic/knowledge transition problem which is particularly important for knowledge-grounded conversations. Merely injecting new knowledge into a generated response does not necessarily improve the quality of a conversation. The knowledge for the next response needs to be carefully chosen so that it is coherent to the historical context of the ongoing conversation, while at the same time sufficiently diversified so as to further engage the user.

However, most of the existing methods (Lian et al., 2019; Kim et al., 2020; Zheng et al., 2020b; Zhao et al., 2020) directly rely on the dialog context to select the next knowledge, which finally lead to spurious correlations. For example, some methods just choose a knowledge that have appeared in dialog history. As shown in the top example in Figure 1, the chosen knowledge *blue* was a repetition of the main topic of the previous turn. A recent

work (Zheng et al., 2020a) considered the differences between the knowledge used in two consecutive turns. However, focusing only on the knowledge differences for knowledge selection could sacrifice the dialog coherence. As shown in the bottom example in Figure 1, *huskies as pet* introduced a huge topic shift that is incoherent with the dialog context of *working dogs*. It is thus important to simultaneously consider both the contextual relevance and the topic development in selecting the next knowledge for response generation.

We propose a new method called C2T2 (**C**oherent and **C**aptivating **T**opic **T**ransition) to select the next knowledge for response generation by simultaneously considering historical contextual coherence and sequential topic development. Specifically, we design transition-aware features to consider both the adherence and diversity of the candidate knowledge for topic transitions. At the same time, we adopt the KL-divergence based shifting loss as a shifting constraint to manage the knowledge variance between turns.

However, making a final decision on a suitable target knowledge could still be tricky even with the above considerations. It is important to also take into consideration the potential relations among the various candidate knowledge when selecting an appropriate next knowledge. Instead of purely doing matching between the (vertical) dialog context and candidate knowledge, we compare all the candidate knowledge in a graph structure by proposing a variant of the PtrNet (Vinyals et al., 2015), named as the Interactive Knowledge Inference module, to also take into consideration the horizontal comparisons of all the target knowledge candidates, in order to reason under all the factors related to dialog topic transition, such as context and transition features, before making the final selection/inference.

In summary, our contributions are three folds. First, we designed novel topic transition features to coherently select appropriate knowledge for creating engaging transitions by effectively managing the historical contextual coherence and sequential topic development under a topic shifting constraint. Second, instead of simply matching between context and knowledge, we proposed an interactive knowledge inference module to model relations between the previously mentioned knowledge and target knowledge candidates, and select the target knowledge based on a comprehensive comparison of all the candidate knowledge in a pointer network. Finally, we showed that our C2T2 outperformed the state of the art methods based on evaluation experiments on two public benchmarks, with gains over 5% and 18% on knowledge selection accuracy in seen and unseen scenarios respectively.

2 Related Work

Knowledge-Grounded Conversation. Knowledge grounded conversations enrich dialog content with external knowledge. The knowledge can be structure-based (Moon et al., 2019; Xu et al., 2020) or document-based (Dinan et al., 2019). The latter is the main topic of this paper. Recently, (Dinan et al., 2019) presented a benchmark where knowledge is explicitly labelled for each conversation turn to explore knowledge selection logic given the dialog context. Knowledge grounded conversations can then be decomposed into two sub-tasks, namely, knowledge selection and response generation. Existing methods mostly deal with knowledge selection by directly matching the dialog context and the potential next knowledge. (Lian et al., 2019) used the posterior knowledge distribution given response to calibrate the context-knowledge mapping. (Zheng et al., 2020b) further exploited context-knowledge and response-knowledge relations in both word and sentence level. With the selected knowledge and dialog context as input, these models then adopt common language decoders (Vaswani et al., 2017; Radford et al., 2019) to generate responses. There are also studies (Zhao et al., 2020; Lin et al., 2020; Rashkin et al., 2021) addressing the response generation with chosen knowledge. In this paper, we focus on the knowledge selection task and adopt GPT-2 (Radford et al., 2019) as our response generator.

Topic Transition Modelling. Topic transition modelling is dealt as a knowledge/topic selection task in knowledge-grounded conversations, in other words, learning or modelling the transition logic from dialog history (including historical conversation and knowledge) to the next knowledge. (Kim et al., 2020) learned historical knowledge sequences by latent variables. (Zheng et al., 2020a) looked at the difference between historical knowledge and next one. (Meng et al., 2020) designed the knowledge tracker and shifter to model knowledge interactions between turns. (Zhan et al., 2021b) extracted topic labels for knowledge to reduce sequential transition noises. However, the challenge of maintaining well-balanced coherence

and knowledge diversity in knowledge-grounded conversations, which was also highlighted in traditional dialog system studies (Li et al., 2016, 2017), was not addressed by these methods. In this paper, we propose an effective way to address the challenge of document-based dialog knowledge selection task to generate responses that are coherent to the dialog context while introducing suitably new knowledge to keep user engaged.

3 Method

3.1 Task Formulation

At each turn t in a knowledge-grounded conversation between a user and an agent (the chatbot), we need to predict the agent's next response r_t given the dialog context $U_t = \{u_{t-l}, r_{t-l}, ..., r_{t-1}, u_t\}$ and the knowledge pool $D_t = \{d_1^t, ..., d_M^t\}$, where l is the number of turns of the context, u_i, r_i are utterances from the user and the agent, respectively. M is the number of the relevant knowledge entries. The two steps of knowledge-grounded conversation namely, knowledge selection and response generation can then be formulated as $P(d_m^t|U_t, D_t)$ and $P(r_t|U_t, d_m^t)$.

3.2 Overview of C2T2

The overall architecture of our proposed C2T2 is shown in Figure 2. It is composed of Sentence Encoder, Topic Transition Modelling, Knowledge Shifting Constraint, Interactive Knowledge Inference and Decoder.

3.3 Sentence Encoder

Similar to (Zhao et al., 2020), we adopt BERT (Devlin et al., 2018) to obtain the embedding for each utterance in the dialog context and each knowledge candidate. Specifically, the utterances in the dialog context are concatenated to be $C_t = [u_{t-l}; r_{t-l}; ...; r_{t-1}; u_t]$. C_t is then combined with each knowledge separately to form paired inputs to BERT, the set of paired inputs I is:

$$I = \{[\text{CLS}]\ C_t\ [\text{SEP}]\ d_m^t\}_{m=1}^M \quad (1)$$

As shown in Figure 2, these paired inputs are fed into BERT to model the correlation between the context and each knowledge candidates and yield their representations. After BERT encoding, we get the hidden state for each token. The hidden state of the special token [CLS], $\mathbf{k}_{m,[\text{CLS}]}^t \in \mathbb{R}^d$, represents the *context-aware knowledge embedding*, where d is the vector dimension. $\mathbf{k}_{m,[\text{CLS}]}^t$ not only incorporates the information of the context C_t and the knowledge d_m^t but also embodies the semantic relations between them, for example, entailment or transitional relation. Additionally, we compute the representations of the context $\mathbf{c}_m^t \in \mathbb{R}^d$ and knowledge $\mathbf{k}_m^t \in \mathbb{R}^d$, by the averaging all the token hidden states in their positions, this process for each pair input I_m is denoted as:

$$\tilde{\mathbf{k}}_m^t, \mathbf{k}_{m,[\text{CLS}]}^t, \mathbf{c}_m^t, \mathbf{k}_m^t = \text{BERT}(I_m) \quad (2)$$

where $m \in [1, ..., M]$ and $\tilde{\mathbf{k}}_m^t$ is the pooled output for the whole input. Besides, we further aggregate all the M context representations $\mathbf{c}_{m1}^{t\ M}$ by attention mechanism, which is formulated as:

$$\mathbf{h}_m^t = \tanh(W_c \mathbf{c}_m^t)$$
$$\alpha_m^t = \frac{\exp(V_c \mathbf{h}_m^t)}{\sum_{i=1}^M \exp(V_c \mathbf{h}_i^t)}, \quad \mathbf{c}^t = \sum_{i=1}^M \alpha_i^t \mathbf{h}_i^t \quad (3)$$

$W_c \in \mathbb{R}^{d \times d}$ and $V_c \in \mathbb{R}^d$ are trainable weights.

3.4 Topic Transition Modelling

We propose a novel transition modelling method that takes into account both dialog coherence and topic development to obtain a transition-aware knowledge representation for a candidate knowledge that effectively capture its historical contextual coherence and sequential topic development.

Context Knowledge Entailment. The context-aware knowledge embedding $\mathbf{k}_{m,[\text{CLS}]}^t$ is the hidden state of the token [CLS] in the last layer in BERT, which encodes the coherence information of sentence pairs thanks to BERT's Next Sentence Prediction pre-training scheme (Devlin et al., 2018). We employ a single fully connected layer with $\tanh$ activation to get the coherence features $\mathbf{v}_{t,m}^{coh} \in \mathbb{R}^2$. This entailing feature $\mathbf{v}_{t,m}^{coh}$ will be integrated with other features to form a comprehensive transition-aware knowledge representation for each candidate knowledge d_m^t.

Sequential Knowledge Development. To model the knowledge changes in sequential topic development, we compute the knowledge difference between the historical knowledge and each candidate knowledge. We first obtain the context-aware knowledge embedding from BERT for the ground-truth knowledge at turn $t-1$, denoted as $\mathbf{k}_{gt,[\text{CLS}]}^{t-1}$. Inspired by (Chen et al., 2017), we apply the cross operator $f(\mathbf{u}, \mathbf{v}) = [\mathbf{u} - \mathbf{v}; \mathbf{u} \odot \mathbf{v}]$, combined difference and element-wise product, to model high-order interaction between the hidden state of the

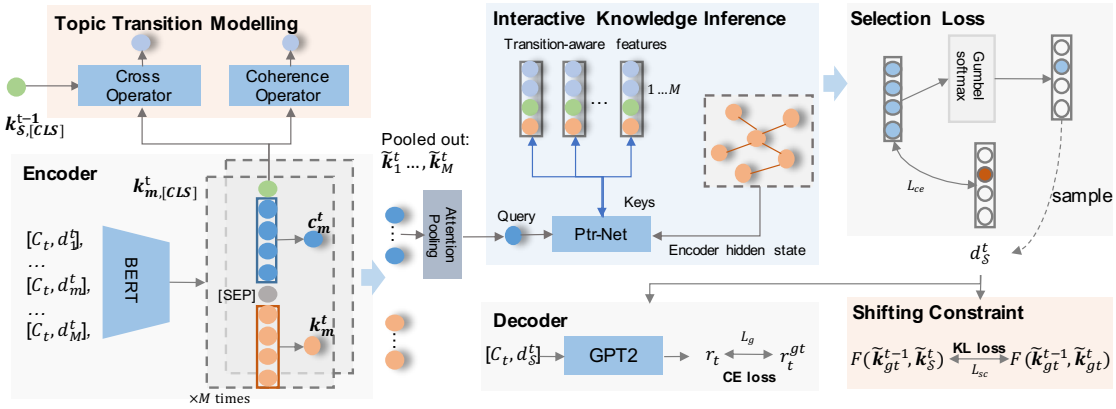

Figure 2: The architecture of C2T2. The left gray part is the outputs of each paired context-knowledge $[C_t; d_m^t]$ from the BERT Encoder. The Topic Transition Modelling and Shifting Constraints (in orange) module control the topic transitions. The Interactive Knowledge Inference (in blue) outputs the selected knowledge d_{S}^t. The right gray part compute the selection loss and ouput the selected knowledge d_{S}^t by gumbel-softmax. The below gray part is the GPT-2 based decoder with context C_t and d_{S}^t as input.

last chosen knowledge $\mathbf{k}_{gt,[\mathrm{CLS}]}^{t-1}$, and each candidate knowledge $\mathbf{k}_{m,[\mathrm{CLS}]}^{t}$, denoted as:

$$\begin{aligned} \mathbf{q}_m^t &= f(\mathbf{k}_{gt,[\mathrm{CLS}]}^{t-1}, \mathbf{k}_{m,[\mathrm{CLS}]}^{t}) \\ \mathbf{v}_{t,m}^{cro} &= \tanh(FC(\mathbf{q}_m^t)), \quad m \in [1, ..., M] \end{aligned} \tag{4}$$

We set $\mathbf{v}_{t,m}^{cro}$ to zero vector where there is no last knowledge, for example, at the first turn of a conversation. The cross operator captures the transition associations between the last knowledge and the candidate knowledge, which are then fed into feed-forward neural networks activated with $\tanh$.

Transition-aware Knowledge Representation. The context knowledge entailment feature $\mathbf{v}_{t,m}^{coh}$, sequential knowledge difference $\mathbf{v}_{t,m}^{cro}$, context-aware knowledge embedding $\mathbf{k}_{m,[\mathrm{CLS}]}^{t}$ and knowledge embedding $\mathbf{k}_m^t$, as described above, together form the transition-aware knowledge representations $\mathbf{E}^t$ for M relevant knowledge, which is denoted as:

$$\begin{aligned} \mathbf{e}_m^t &= [\mathbf{v}_{t,m}^{coh}; \mathbf{v}_{t,m}^{cro}; \mathbf{k}_{m,[\mathrm{CLS}]}^{t}; \mathbf{k}_m^t] \\ \mathbf{E}^t &= (\mathbf{e}_1^t, ..., \mathbf{e}_M^t) \end{aligned} \tag{5}$$

where $[;]$ means the concatenation operator along the last dimension of tensor.

3.5 Knowledge Shifting Constraint

We devise a Knowledge Shifting Constraint to control the variance between knowledge in consecutive turns. The constraint is an auxiliary loss in training phase to ensure the variance of transitions as follows. Given the indexes of the ground-truth knowledge at the $t-1$ turn, the ground-truth knowledge and the selected knowledge by Gumbel-Softmax (Jang et al., 2016) at current turn t, we use their pooled outputs from BERT $\tilde{\mathbf{k}}_{gt}^{t-1}$, $\tilde{\mathbf{k}}_{gt}^{t}$ and $\tilde{\mathbf{k}}_{\mathrm{S}}^{t}$ to compute the information variance of two tuples, $\langle \tilde{\mathbf{k}}_{gt}^{t-1}, \tilde{\mathbf{k}}_{\mathrm{S}}^{t} \rangle$ and $\langle \tilde{\mathbf{k}}_{gt}^{t-1}, \tilde{\mathbf{k}}_{gt}^{t} \rangle$. The former measures the difference between the current selection and the previous knowledge, while the latter computes the difference between the ground-truth selection and the previous knowledge. These two distributions should be close to each other to keep the variance between the former knowledge and current knowledge. Therefore, we adopt the Kullback-Leibler divergence to narrow the difference of these two distributions, which is made as an auxiliary loss, L_{sc}, denoted as:

$$L_{sc} = D_{KL}(F(\tilde{\mathbf{k}}_{gt}^{t-1}, \tilde{\mathbf{k}}_{\mathrm{S}}^{t}) \parallel F(\tilde{\mathbf{k}}_{gt}^{t-1}, \tilde{\mathbf{k}}_{gt}^{t})) \tag{6}$$

We define the variance measure function as:

$$F(\mathbf{u}, \mathbf{v}) = \mathrm{log\_softmax}([(\mathbf{u}-\mathbf{v})^2; \mathbf{u} \odot \mathbf{v}]) \tag{7}$$

3.6 Interactive Knowledge Inference

Instead of independently matching context with each candidate knowledge, we adopt a variant of Ptr-Net (Vinyals et al., 2015) to compare all the knowledge candidates to select the target knowledge by comprehensively considering dialog context, knowledge candidates and their relations in a graph structure.

As Ptr-Net encodes the input in a sequence structure which is not applicable in our situation, we introduce a multi-head graph attention network (Veličković et al., 2018) to encode the associations of all the candidate knowledge within a graph structure, with each knowledge as a node. The graph structure $\mathcal{G}$ is constructed based on the text-similarity (tf-idf) of the candidate knowledge

sentences. Each node is initialized by the knowledge representation $\mathbf{k}_m^t$ from the BERT encoder. The output embedding of all nodes is then fed into an average pooling layer to obtain the graph representation $\mathbf{h}^t$, capturing the relationship and semantic features of all knowledge candidates related to the dialog context. The process is formulated as:

$$\mathbf{h}^t = \text{avgpool}(\text{GAT}([\mathbf{k}_m^t]_1^M, \mathcal{G})) \tag{8}$$

Ptr-Net is a sequence decoder and we set the decoding length to 1 in this task to only choose one knowledge for each turn. In a summary, with dialog context $\mathbf{c}^t$ as query, transition-aware knowledge representations $\mathbf{E}^t$ as keys, and knowledge interactive representation $\mathbf{h}^t$ as the encoder hidden state, the Ptr-Net decodes knowledge as:

$$\begin{aligned}
\hat{\mathbf{h}}^t &= \text{LSTMCell}(\mathbf{c}^t, \mathbf{h}^t) \\
\hat{\alpha}_m^t &= \mathbf{v}^\top \tanh(W_e \mathbf{e}_m^t + W_h \hat{\mathbf{h}}^t + \mathbf{b}) \\
P(d_m^t | U_t, D_t) &= \frac{\exp(\hat{\alpha}_m^t)}{\sum_{i=1}^M \exp(\hat{\alpha}_i^t)}
\end{aligned} \tag{9}$$

$W_e, W_h \in \mathbb{R}^{d\times d}$,$\mathbf{v}, \mathbf{b} \in \mathbb{R}^d$ are trainable weights.

In a knowledge-grounded conversation dataset, the turns of all the conversations form the samples $\mathcal{D} = \{(U_i, D_i, r_i)\}_1^N$ where U_i, $D_i = \{d_m^i\}_1^M$ and r_i are the dialog context, candidate knowledge and response respectively. The knowledge selector model is trained by minimizing the loss function L_{cls} on $\mathcal{D}$ as follows:

$$\begin{aligned}
L_{cls} &= L_{ce} + \lambda L_{sc} \\
L_{ce} &= -\frac{1}{N}\sum_{i=1}^N \hat{\mathbf{y}}_i \log(p(d_i | U_i, D_i))
\end{aligned} \tag{10}$$

where $\hat{\mathbf{y}}_i$ denotes the one-hot vector indicating the ground-truth knowledge for data sample i. $p(d_i|U_i, D_i)$ denotes the probability distribution over the candidate knowledge D_i. L_{ce} is a standard cross-entropy loss function for knowledge selection and λ is the coefficient that makes a balance between the two objective functions.

3.7 Response Generation

The response r_i is finally generated by GPT-2 model, given dialog context U_i and the selected knowledge sentence d_{S}^i from the knowledge selector. The GPT-2 model generates a distribution over the vocabulary $\mathcal{V}$ at each decoding position, which is fine-tuned with the cross-entropy loss:

$$\begin{aligned}
p(r_\tau^i | U_i, d_{\mathrm{S}}^i, r_{<\tau}^i) &= \text{GPT-2}(d_{\mathrm{S}}^i, U_i, r_{<\tau}^i) \\
L_g &= -\frac{1}{N}\frac{1}{|r_i|}\sum_{i=1}^N \sum_{\tau=1}^{|r_i|} \hat{\mathbf{y}}_\tau^i \log p(r_\tau^i | U_i, d_{\mathrm{S}}^i, r_{<\tau}^i)
\end{aligned} \tag{11}$$

where $\hat{\mathbf{y}}_\tau^i \in \mathbb{R}^{|\mathcal{V}|}$ is the one-hot vector indicating the ground-truth word at position τ of response. $p(r_\tau^i | U_i, d_{\mathrm{S}}^i, r_{<\tau}^i)$ is the probability distribution over the vocabulary at position τ.

	Seen			Unseen		
Model	**ACC**	**uni-F1**	**BLEU1/2**	**ACC**	**uni-F1**	**BLEU1/2**
TMN	23.2	17.7	-	12.2	14.4	-
PostKS	23.4	18.1	-	9.4	13.5	-
BERT+PoKS	25.5	17.8	-	14.1	13.4	-
KIC	-	18.9	17.3/10.5	-	17.3	16.5/9.5
PIPM	27.8	-	-	19.4	-	-
DukeNet	26.4	19.3	18.0/7.5	19.6	17.1	16.3/6.0
SLKS	26.8	19.3	18.9/10.9	18.3	16.1	17.3/8.0
SLKS+GPT2	26.8	20.5	18.8/9.9	18.3	17.7	16.4/7.7
DiffKS+GPT2	25.6	21.1	18.8/10.2	18.6	18.6	17.4/8.6
KnowledGPT	28.0	21.9	19.5/10.8	25.4	19.6	17.7/9.1
CoLV	30.1	-	-	18.9	-	-
C2T2	**31.7**	**22.4**	**20.2/11.4**	**30.1**	**21.1**	**19.2/10.4**

Table 1: Experimental Evaluation results on WoW dataset. ACC is accuracy for knowledge selection. ROUGE1/2 scores can be found in Appendix D

4 Experiments

4.1 Datasets

We evaluated our model on two commonly used public benchmark datasets for the knowledge grounded dialog system, Wizard of Wikipedia (WoW) (Dinan et al., 2019) and Holl-E (Moghe et al., 2018). WoW consists of 18,430/1,948/965/968 dialogs for train/valid/test_seen/test_unseen split. Each dialog is constructed in wizard-apprentice style, and the wizard tries to inform the other person about the Wikipedia topic. Holl-E contains conversations about movies, and each response is based on some background documents. Following (Kim et al., 2020), we adopted the 7,211/930/913 dialogs splits for train/valid/test for Holl-E.

4.2 Evaluation Metrics

Following previous work, we evaluate the two subtasks of knowledge selection and response generation. We use accuracy (**ACC**) to measure the performance of knowledge selection. We further design two metrics Adhesion (**Adh.**) and Diversity (**Div.**) to evaluate the knowledge coherence

Model	Seen			Unseen		
	Div.	Adh.	QDCE	Div.	Adh.	QDCE
DukeNet	16.4	51.3	3.205	10.6	45.8	3.215
SLKS	13.2	54.5	3.298	6.9	39.9	3.294
DiffKS	15.2	49.7	3.211	11.1	36.0	3.204
KnowledGPT	16.5	49.8	3.351	11.5	56.5	3.363
C2T2	**20.4**	**57.9**	**3.445**	**18.0**	**57.7**	**3.459**

Table 2: Knowledge transition and dialog coherence evaluations on WoW dataset. Adh. and Div. are fine-grained knowledge selection accuracy of dialog turns where knowledge is the same as or different from the last knowledge. QDCE denotes QuantiDCE (Ye et al., 2021a), which measures coherence between dialog context and the generated response.

and diversity, which are defined as the ratio of the correct knowledge selection in dialog turns where knowledge had remained the same as, or different from, the knowledge used in last turn, respectively.

For the response generation, **uni-gram F1**, **BLEU1/2**(1/2 means uni-gram and bi-gram), **ROUGE1/2** are used to automatically measure the similarity between generated response and the ground-truth in token and phrase level. For dialog-level coherence, we adopted a state-of-the-art well-trained model **QuantiDCE** (Ye et al., 2021b) in Automatic dialog Coherence Evaluation task to measure the generated dialog coherence.

4.3 Baselines

We compare our C2T2 with most of the existing approaches for knowledge-grounded conversation.
TMN (Dinan et al., 2019). The transformer with memory network is the baseline model along with the release of the Wizard of Wikipedia.
PostKS (Lian et al., 2019) learns the knowledge selection with help of posterior distribution and the advanced version **BERT+PoKS** (Dinan et al., 2019) with BERT as encoder.
SLKS (Kim et al., 2020) first sequentially models knowledge selection and decodes response with the Transformer with copy mechanism.
PIPM (Chen et al., 2020) improves SLKS by learning complement posterior knowledge information which is missing in inference stage for SLKS.
DukeNet (Meng et al., 2020) models knowledge tracking and knowledge shifting as dual tasks.
KIC (Lin et al., 2020) deals with response generation by copying words from knowledge with pointer network.
CoLV (Zhan et al., 2021a) uses a collaborative latent variable model to integrate knowledge selection and knowledge-aware response generation.
KnowledGPT (Zhao et al., 2020) compatibly combines pre-trained language models for knowledge selection and response generation.

For fairer comparison, we replaced some baselines with the same (more powerful) response generator GPT-2 as ours, such as SLKS (Kim et al., 2020) and DiffKS (Zheng et al., 2020a). This resulted in another two baselines **SLKS+GPT2** and **DiffKS+GPT2**. Another recent work DIALKI (Wu et al., 2021) regarded knowledge selection as knowledge identification in a long document, which exploits extra knowledge position information in corresponding wiki articles. It is unfair to compare this method with all previous work such as SLKS and CoLV. In fact, by adding this extra position information, our method also outperformed DIALKI, reaching 34.5/35.6 in terms of ACC on Test Seen and Unseen of WoW.

4.4 Implementation Details

Most of the code were based on Pytorch (Paszke et al., 2019). For the implementation of BERT(110M) and GPT-2(117M), we used the package from the Huggingface Transformers[1] (Wolf et al., 2020). Adam (Kingma and Ba, 2015) was the optimizer for both knowledge selector and response generator. The training batch size and initial learning rate for BERT and GPT-2 were 4 and 32, 1e-5 and 5e-5, respectively. In the knowledge selector, the learning rate for modules other than BERT is 1e-4. A linear scheduler with a warm-up for the learning rate was used in knowledge selection. For the response generation, we gradually reduced the ratio of ground-truth knowledge in generation training following (Zhao et al., 2020). It took around 5 and 10 epochs to achieve the reported performance in knowledge selection and generation. We set the balance coefficients λ and μ for L_{sc} and L_g to 0.5 and 2. We will release all the codes and hyper-parameters setting for re-production.

4.5 Analysis

In this part, we mainly analyze our experiments from four research questions.
Q1. Is C2T2 able to perform well on knowledge selection and response generation? The experimental evaluation results on WoW and Holl-E are shown in Table 1, Table 2 and Table 3. As illustrated in Table 1 and Table 2, our C2T2 model

[1]https://github.com/huggingface/transformers

Model	ACC	uni-F1	BLEU1/2	ROUGE1/2
SLKS	29.2	29.8	28.0/22.2	31.3/23.2
DiffKS+GPT2	33.5	31.9	31.2/26.9	33.9/24.7
PIPM	30.7	-	-	30.8/24.0
DukeNet	30.0	30.6	30.1/22.5	**36.5**/23.0
CoLV	32.7	-	-	32.0/**25.8**
C2T2	**37.7**	**32.9**	**31.8/28.0**	34.8/25.6

Table 3: Evaluation results on Holl-E dataset. The best results are highlighted with **bold**.

Model	Seen ACC	Seen Div.	Seen Adh.	Seen uni-F1	Unseen ACC	Unseen Div.	Unseen Adh.	Unseen uni-F1
C2T2	31.7	20.4	57.9	22.4	30.1	18.0	57.7	21.1
w/o ShiftLoss	30.6	19.5	56.0	22.0	29.5	17.0	58.6	21.3
w/o CrossOpt	30.2	18.9	56.4	21.9	28.4	14.8	59.0	20.6
w/o CoherOpt	30.0	19.3	54.8	21.9	28.8	16.6	56.8	20.8
w/o PointerNet	29.7	18.6	55.5	21.8	29.6	17.2	57.7	21.1

Table 4: Ablation test results on WoW dataset. Almost all parts contribute to the C2T2 final performance in the four metrics. One exception is on Adh. Adh.increases after removing Shift loss or Cross Operator.

outperformed all the baselines in terms of all metrics on the Seen and Unseen test sets of WoW.

For knowledge selection, C2T2 significantly outperformed the very recent work CoLV in ACC by 1.6% and 11.2% on Test Seen and Test Unseen respectively. Even compared with knowledGPT, which has the best comprehensive performance on both test sets, our method also improved by 3.7% and 4.7%, reaching 31.7/30.1. It is particularly worth noting that our method outperformed five strong baselines SLKS, DiffKS, DukeNet, PIPM and CoLV on the more difficult WoW Test Unseen dataset by 11.8%, 11.5%, 10.5%, 10.7% and 11.2%, proving that C2T2 model could capture the more generalized patterns of topic transitions.

For response generation, C2T2 also achieved the best results on all automatic metrics(F1, BLEU1/2, ROUGE1/2 and QuantiDCE), which showed the dialog response from C2T2 performs best in fluency, relevance and coherence. Moreover, compared with SLKS+GPT2, DiffKS+GPT2 and KnowledGPT which used the same GPT2 as generator, our model also significantly outperformed these strong baseline models, confirming that higher knowledge selection accuracy do lead better generation. Note that KnowledGP also used the same encoder and decoder, BERT and GPT2 as C2T2. C2T2 still outperformed the KnowledGPT even though the latter adopted more sophisticated and costly training strategies (the reinforcement step and the curriculum step).

Similar results were also observed on Holl-E in Table 3. C2T2 achieved significant performance gains on all the metrics compared to other baselines, showing the highest accuracy in knowledge selection with margins of 8.5%, 4.2%, 7.7%, 7% and 5% with respect to five strong baselines SLKS, DiffKS, DukeNet, PIPM and CoLV.

Q2. Whether C2T2 improves topic coherence and knowledge diversity? As shown in Table 2, compared with four strongly baselines, C2T2 achieved highest score on Div. and Adh., with margins of $4\%/7.4\%$, $7.2\%/11.1\%$, $5.2\%/6.9\%$, and $3.9\%/6.5\%$ on two test sets of WoW, demonstrating that C2T2 indeed improved both dialog topic adhesion and knowledge diversity. Moreover, the result of QuantiDCE, which measures the overall dialog coherence from the generated conversations, further validates that our C2T2 is able to capture better dialog transition logic.

Q3. Whether each module contributes to of C2T2 Performance? We conducted a series of ablation experiments on the WoW dataset. Four variants were designed for ablation study as follows: (1) *w/o ShiftLoss*: removing the Shifting Constraint Loss; (2) *w/o CrossOpt*: cutting the cross operation between candidate knowledge and previous selected knowledge ;(3) *w/o CoherOpt* removing coherence operator; (4) *w/o PointerNet*: replacing Interactive Knowledge Inference with a simpler knowledge selection module, where knowledge selection distribution is defined as the attention scores between the context hidden $\mathbf{c}_t$ and each transition-aware knowledge representations $\mathbf{E}_t$. Almost all the results of these variants, as shown in Table 4, exhibited performance drops on knowledge selection and response generation, showing their contribution to our model's generalisation ability.

Q4. Which parts of C2T2 improve coherence and diversity? In Table 4, we observe in unseen column that after removing cross operator or shift loss, the performance of Adhesion improves while Diversity declines compared to C2T2, which indicates that these two modules do advocate topic change and suppress same topic. For *w/o CoherOpt*, the Adh. performance drops the most after removing coherence operator, showing that coherence operator promotes knowledge adhesion.

4.6 Human Evaluation

To make up the the shortcomings of automatic experimental evaluations, we also conducted human evaluation. The human evaluation criteria consist of two parts, naturalness and appropriateness.

Methods	Naturalness Win	Lose	κ	Appropriateness Win	Lose	κ
Wow Seen						
C2T2 vs. SLKS	85	7	0.43	70	10	0.43
C2T2 vs. KnowledGPT	34	18	0.40	30	12	0.38
C2T2 vs. DukeNet	79	11	0.50	74	11	0.43
Wow Unseen						
C2T2 vs. SLKS	78	6	0.32	65	12	0.31
C2T2 vs. KnowledGPT	28	17	0.31	24	18	0.32
C2T2 vs. DukeNet	69	11	0.45	61	11	0.32

Table 5: Human evaluation on WoW dataset. *Win* and *Lose* are the percentage of C2T2 wins or loses comparing to other methods, the remain part of *tie* is omitted.

The former emphasizes the readability and fluency of the sentence itself, while the latter highlights whether appropriate knowledge information is used in the response given the context of the conversation. We randomly selected 300 samples from seen and unseen test set of WoW, and three curators evaluated each sample on the Amazon Mechanical Turk according to the criteria. The results are shown in Table 5. Our method significantly outperformed SLKS and DukeNet in both criteria. For knowledGPT, although C2T2 and knowledGPT both used GPT-2 as the generator, C2T2 still performed better. We also computed the FLeiss' Kappa (Fleiss, 1971) to measure the agreement of all the curators on each sample. Due to more than 20 people participated in this evaluation study, we have a moderate fleiss' kappa value within the range 0.3-0.5.

4.7 Case Study

In Figure 3 we visualize the knowledge selection and response generation results of different models. In turn *T2*, KnowledGPT introduced incoherent knowledge, *Chanel s.a...*, which has little relation with last topic *fragrance chanel5*. DukeNet failed to convey new information but fell into the spurious correlation of repeating history due to the lack of constraints on transition. SLKS produced an unreadable response. In *T3*, KnowledGPT and DukeNet repeated the User's utterance in *T3* and the Gold Bot's utterance in *T2*, respectively. In comparison, our C2T2 can introduce more informative knowledge. Even in this case where the knowledge selected was not consistent with the ground-truth, C2T2 also selected coherent knowledge, for example, **(k4)** in *T3*.

5 Conclusion

Effective topic transition modelling is critical for a coherent and captivating conversation. In this paper, we proposed a new method named C2T2 that simultaneously considered topic coherence and knowledge diversity for topic transitions in knowledge grounded conversations. C2T2 endows knowledge-grounded dialog systems the ability to coherently select appropriate knowledge for creating transitions. Instead of purely matching between dialog context and candidate knowledge, our C2T2 comprehensively compare the previously mentioned knowledge and target knowledge candidates with factors such as context and transition features as well as knowledge relations before selecting the next knowledge for response generation. Our C2T2 method achieved the new state-of-the-art performance on both Wizard of Wikipedia (Dinan et al., 2019) and Holl-E (Moghe et al., 2018), in particular, made significant progress in the knowledge selection task. However, measuring knowledge coherence and the appropriateness of diversity is a non-trivial problem. In this paper we simply use accuracy of the knowledge change and knowledge adhesion to reflect these two properties, which could be replaced by more well-defined metrics.

Topic: Coco Chanel .		
k1: Along with Paul Poiret, Chanel was credited in the post-world war I era with liberating women from the constraints of the "corseted silhouette" and popularizing a sporty, casual chic ...		
k2: Chanel s.a. is a high fashion house that specializes in haute couture and ready-to-wear clothes, luxury goods and fashion accessories.		
k3: Chanel no . 5, traditionally , fragrance worn by women had adhered to two basic categories : respectable women favored the pure essence of a single		
k4: A prolific fashion creator, Chanel extended her influence beyond couture clothing, realising her design aesthetic in jewellery, handbags, and fragrance.		
k5: As a fashion designer, Coco Chanel catered to women's taste for elegance in dress, with blouses..... of simple design....		
k6: In her youth , Gabrielle Chanel gained the nickname " coco " from her time as a chanteuse .		
T1	***User**: Coco Chanel .* ***Gold Bot**: I learned that the founder of the chanel brand ' s (1) first name is coco. are you familiar with the brand?*	
T2	***User**: Yes, I wear her fragrance chanel #5 for the last two decades .* ***Gold Bot**: I also did not realize how far back the Chanel brand goes. post wwi, coco along with paul poiret (2) made popular a look for women that was more sporty and chic.*	
	C2T2	**(k1)** I know that Paul Poiret was credited with her inventing the sporty, casual chic.
	Knowled GPT	**(k2)** Chanel s.a. is a high fashion house that specializes in haute couture and ready-to-wear clothes, luxury goods and
	SLKS	**(k3)** Yes , she is a great brand . she is a respectable brand
	DukeNet	**(k1)** I love coco chanel. she is the founder of the chanel brand (1).
T3	***User:** Yes Coco Chanel was a french fashion designer(3) and a business woman(4) who became an icon(5).* ***Gold Bot:** Its fascinating her rise in the fashion world. as a matter of fact, she is the only fashion designer to be listed in the time magazine*	
	C2T2	**(k4)** Yes, she was. she also created handbags and fragrance. she also made jewellery and handbags.
	Knowled GPT	**(k5)** Yes, she was a fashion designer(3) who was a fashion designer who was a business woman(4) who was a fashion icon(5).
	SLKS	**(k6)** I know ! in her youth gabrielle chanel was a business woman (4).
	DukeNet	**(k1)** Chanel was credited with paul poiret with paul poiret(2)

Figure 3: A Complete conversation with three turns. Turn T2 and T3 show the generation comparisons of different methods. Blue words indicate repetition, attached with a number pointing to the dialogue history text with the same number. Red words show incoherence. Sentences with prefix ***User*** and ***Gold Bot*** are user utterances and ground-truth responses in this turn.

6 Ethical Impact

Our work study towards coherent and captivating topic transitions method in knowledge-grounded dialog systems. Both WoW and Holl-E we evaluate our model on were privacy filtered and content moderated by the original authors (Dinan et al., 2019; Moghe et al., 2018). Further research work in knowledge selection of the knowledge-grounded dialog systems based on our work or findings is encouraged. After we release source codes of the whole system, people can build open-domain intelligent dialogue bots based on our code and Internet vast data to serve lots of users. On the other hand, knowledge selection techniques in knowledge-grounded conversations may also be used to build conversational bots that serve illegal purposes or select and generate inappropriate or even harmful content.

For the human evaluation part, after consulting the ethics review board in our lab, we recruited workers from Amazon Mechanical Turk to do the human evaluation. All the 25 workers are from United States, and each data sample costs 0.4 dollars. The protocol of using this data is defined by the agreement on Amazon Mechanical Turk and it is fine to use for research use.

References

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 733–743, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI International Joint Conference on Artificial Intelligence*, page 5081.

Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 41–52.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1151–1160.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over

knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in Neural Information Processing Systems*, 28:2692–2700.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. Dialki: Knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1835–1845, Online. Association for Computational Linguistics.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021a. Towards quantifiable dialogue coherence evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2718–2729, Online. Association for Computational Linguistics.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021b. Towards quantifiable dialogue coherence evaluation. In *Proceedings of the 2021 Conference on Association for Computational Linguistics (ACL)*.

Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021a. Colv: A collaborative latent variable model for knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2250–2261.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021b. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020a. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 115–125.

Wen Zheng, Natasa Milic-Frayling, and Ke Zhou. 2020b. Approximation of response knowledge retrieval in knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3581–3591.

A Implementation Details

We run our experiments on 2 Geforce RTX 3090 GPUs. For selection, each knowledge is concatenated with dialogue context as the input to the BERT model. The max length of this concatenated input is 96. The hidden size of the attention aggregation module for context embedding is the same as its input size, which is also the same as

Model	Seen ROUGE 1/2	Unseen ROUGE 1/2
PIPM	19.3/7.4	17.6/5.5
DukeNet	**25.2**/6.8	23.3/5.3
SLKS	21.1/7.0	18.2/15.9
SLKS+GPT2	23.4/7.4	20.3/5.2
DiffKS+GPT2	23.9/7.9	21.3/5.8
KnowledGPT	24.7/8.5	22.3/6.5
CoLV	20.6/7.9	19.7/6.3
C2T2	**25.2/8.9**	**23.7/7.6**

Table 6: ROUGE scores for response generation

the output size of BERT. The input size of GAT is 768, and the headers of GAT are 8, so we set the hidden size to 96 to keep the output size of GAT the same as the input. A position-wise feed-forward layer (Vaswani et al., 2017) is adopted after the GAT with the hidden size of 2048. The dropout rate for both GAT and feed-forward layers are 0.5 and 0.1, respectively. The hidden size for the pointer-based reasoning module is the same as the input size, 768 for Bert embedding. In training, the temperature of Gumbel-softmax is fixed to 1. For generation, the knowledge and dialogue context are concatenated together as input to GPT-2. The total concatenated input's max length was 256, and the max length for each sentence of knowledge and history was set as 64. As for the training, we first use the ground-truth knowledge to fine-tune the GPT-2 decoder, and the ratio r of using ground-truth in generation will decay with training step s in the rate λ, which can be formulated as $r = 1/e^{s\lambda}$. We set $\lambda = 1e-5$ following (Zhao et al., 2020). This process facilitates generation with the right knowledge provided and has been proved to generate better responses. As for other baselines, all the results are based on the code they provided.

B Human Evaluation Details

We published the evaluation task on Amazon Mechanical Turk and required only native speakers to evaluate generated responses. Precisely, each evaluation sample consists of the dialogue context, four responses from ours, KnowledGPT, SLKS and DukeNet and is evaluated by three different people. First, these curators were given the instructions for criterion definitions and evaluation steps. Then they will go through the dialogue context, compare each pair of responses, and choose a better one from naturalness and appropriateness.

Methods	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
WoW Seen					
PostKS++	56.8	15.6	9.6	6.2	4.1
SLKS	57.4	18.4	10.1	8.9	5.4
DiffKS$_{Fus}$	57.4	22.5	12.8	9.8	7.4
DiffKS$_{Dis}$	56.6	21.5	11.2	10.2	7.9
KnowledGPT	**60.2**	23.6	16.2	12.4	11.2
C2T2	**60.2**	**28.2**	**21.0**	**17.9**	**16.5**
WoW Unseen					
PostKS++	42.8	8.5	4.1	4.8	4.6
SLKS	43.0	6.1	5.2	4.9	5.0
DiffKS$_{Fus}$	40.9	21.2	10.5	7.7	4.6
DiffKS$_{Dis}$	40.2	16.1	10.3	7.7	6.1
KnowledGPT	**63.0**	16.3	12.0	10.3	11.2
C2T2	62.0	**21.7**	**19.9**	**17.7**	**16.1**
Holl-E					
PostKS++	62.8	17.9	18.8	20.0	23.2
SLKS	65.2	18.4	19.2	21.3	19.6
DiffKS$_{Fus}$	**65.8**	22.3	22.1	25.5	25.8
DiffKS$_{Dis}$	63.9	23.0	23.4	26.0	28.3
C2T2	60.5	**32.0**	**32.2**	**31.4**	**31.4**

Table 7: Knowledge selection accuracy over turns.

C Accuracy Over Turns

According to previous work (Kim et al., 2020; Zheng et al., 2020a), the accuracy of knowledge selection will decrease as the turns of dialogue increase. To prove the superiority of our model in deeper turns, we evaluated the accuracy of knowledge selection over turns. As shown in Table 7, our model significantly outperforms almost all the baseline models from 2^{nd} to 5^{th} turns on both benchmarks, which proves that C2T2 can facilitate knowledge selection in deeper turns thanks to simultaneously considering topic relevance and topic development in knowledge grounded conversations. It is worth noting that the accuracy of C2T2 improves more obviously from 4^{th} to 5^{th} turns, with an increase of almost 6%. This is because topic transitions will be more flexible in deeper conversations, and our method can handle this situation better.

D Response Generation Evaluation

We also calculate ROUGE scores on WoW datase. It is shown that our C2T2 perform the best in ROUGE score, indicating that C2T2 can do the best in retrieving similar words and phrases in ground-truth response.

E A Case Study

To give a more intuitive evaluation for the generated conversations, we visualized another conversation about the topic *Halloween*, as shown in Figure 4. In this figure, the k1 to k10 is the list of candidate knowledge, and there are 5 turns of dialogue in total. The post is the utterance from the apprentice user, and *Ground-truth response* is the correct response by the wizard user(the agent we need to model in the KGC task), who tends to convey more knowledge to the apprentice. As described in (Zheng et al., 2020a), for the first round, all the methods have chosen the right knowledge and generated good results. For the *Post 2*, only C2T2 predicted the right knowledge. Even SLKS and DukeNet also provide related responses, but they are less informative. KnowledGPT also introduces new information, but the knowledge shifting is too large from *Halloween* activities to *Halloween origin*. For *Post 3*, both C2T2 and KnowledGPT predicted the right knowledge. In *Post 4*, C2T2, knowledGPT and DukeNet failed to capture the main topic *Michael Myers*, while SLKS provided seemly coherent responses but in fact a repetition of the last utterance from the user. As for *Post 5*, C2T2 transferred topic from *Halloween movies* to *horrible movices*. KnowledGPT, SLKS and DukeNet just chose the knowledge of introducing the movie *Halloween* and failed to focus on the main topic *Halloween movies*. Combining 4 and 5, we can found KnowledGPT tended to choose k6 as long as *movie* and *Halloween* appeared in the post.

Topic: Halloween.	
k1: Halloween or hallowe'en (a contraction of all hallows' evening), also known as all Halloween, all hallows' eve, or all saints' eve, is a celebration observed in a number of countries on 31 October, the eve of the western Christian feast of all hallows' day.	
k2: It begins the three-day observance of allhallowtide, the time in the liturgical year dedicated to remembering the dead, including saints (hallows), martyrs, and all the faithful departed.	
k3: Halloween activities include trick - or - treating (or the related guising) , attending Halloween costume parties , carving pumpkins into jack	
k4: It is widely believed that many Halloween traditions originated from ancient Celtic harvest festivals, particularly the Gaelic festival Samhain; that such festivals may have had pagan roots; and that Samhain itself was Christianized as Halloween by the early church.	
k5: In many parts of the world , the Christian religious observances of all hallows ' eve , including attending church services and lighting candles	
k6: Halloween is a 1978 American slasher film directed and scored by John Carpenter, co-written with producer Debra Hill, and starring Donald pleasance and Jamie Lee Curtis in her film debut.	
k7: A horror film is a movie that seeks to elicit a physiological reaction, such as an elevated heartbeat, through the use of fear and shocking one\u2019s audiences	
k8: Michael Myers is a fictional character from the \"halloween\" series of slasher films.	
k9: He first appears in john carpenter's \"halloween\" (1978) as a young boy who murders his sister and then, fifteen years later, returns home to murder more teenagers	
k10: some believe, however, that Halloween began solely as a Christian holiday, separate from ancient festivals like Samhain.	
***Post 1**: Halloween is the best holiday ever! .*	
Ground-truth response	*(**k1**) Halloween is a celebration observed in a number of countries on 31 October, the eve of the western Christian feast of all hallows' day.*
C2T2	(**k1**) I love it! It's a celebration observed in a number of countries on 31 October, the eve of the western Christian feast
KnowledGPT	(**k1**) I love Halloween, also known as allhalloween, i love it too!
SLKS	(**k1**) I love Halloween ! I love that it is an annual festival !
DukeNet	(**k1**) I love Halloween ! It ' s a great place to visit .
***Post 2**: Yep, i love being able to dress up and watch scary movies.*	
Ground-truth response	*(**k2**) It begins the three-day observance of allhallowtide, the time in the liturgical year dedicated to remembering the dead, including saints(1).*
C2T2	(**k2**) Halloween is the 3 day observance of allhallowtide, the time in the liturgical year dedicated to remembering the dead, including saints
KnowledGPT	(**k4**) I love watching scary movies. Halloween traditions originated from ancient Celtic harvest festivals, particularly the Gaelic festival Samhain.
SLKS	(**k3**) Yes , me too ! trick or treating is so cool !
DukeNet	**(k3)** i love Halloween too ! i like to watch trick or treating pumpkins ..
***Post 3**: I never knew that! What culture is that celebrated in?*	
Ground-truth response	*(**k4**) Its widely believed that many Halloween traditions originated from ancient Celtic harvest festivals, particularly the Gaelic festival*
C2T2	(**k4**) It is believed that Halloween traditions originated from ancient Celtic harvest festivals, particularly the Gaelic festival Samhain; that such festivals
KnowledGPT	(**k4**) It is widely believed that many Halloween traditions originated from ancient Celtic harvest festivals, particularly the Gaelic festival Samhain; that
SLKS	(**k5**) It is celebrated in many parts of the world.
DukeNet	(**k2**) It is celebrated the three - day obserdededetidetide , the time in the liturgical year dedicated to remembering the dead (1).
***Post 4**: That would make sense. what about the Halloween movies with Michael Myers? do you like those?(2)*	
Ground-truth response	(**k8**) *is a fictional character from the "halloween" series of slasher films*
C2T2	(**k6**) I love Halloween, it is a 1978 American slasher film directed and scored by John Carpenter, co-written with producer Debra
KnowledGPT	(**k6**) I do like Halloween, it was directed and scored by John Carpenter, co-written with producer Debra Hill, and starring Donald
SLKS	(**k5**) Yes , i love the Halloween movies with Michael Myers - it ' s so much fun !(2)
DukeNet	(**k9**) I love Michael Carpenter . he was a young boy . he was a young boy .
***Post 5**: Yep, i love those movies. They're my favorites to watch on Halloween.*	
Ground-truth response	*(**k10**) Some believe, however, that Halloween began solely as a Christian holiday, separate from ancient festivals*
C2T2	(**k7**) I love horror films, they elicit a physiological reaction, such as an elevated heartbeat, through the use of fear and shock
KnowledGPT	(**k6**) Is a 1978 American slasher film directed and scored by John Carpenter, co-written with producer Debra Hill, and starring Donald Pleas
SLKS	(**k6**) It was directed by John Carpenter
DukeNet	**(k6)** I love the movie Halloween , it ' s a great American slasher (3).

Figure 4: Blue words indicate repetition, and red words are related to incoherence. The numbers after blue words point to the repeated parts with the same numbers in dialogue history. *Post i* is the dialogue history at turn *i*. At each turn *i*, models try to predict the response of *Ground-truth response*.