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 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 (Coherent and Captivating Topic Transition) 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 PrNet (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 $U_t = \{u_{t-1}, r_{t-1}, ..., r_{t-1}, u_t\}$ and the knowledge pool $D_t = \{d_1^t, ..., d_m^t\}$, where $l$ is the number of turns of the context, $u_t, r_t$ 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 (Zha 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 $C_t = [u_{t-1}; r_{t-1}; ...; r_{t-1}; u_t]$. C$_t$ is then combined with each knowledge separately to form paired inputs to BERT, the set of paired inputs $I$ is:

$$I = \{[\text{CLS}]C_t[\text{SEP}]d_m^t\}_{m=1}^M$$ (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], $k_m^{t, [\text{CLS}]} \in \mathbb{R}^d$, represents the context-aware knowledge embedding, where $d$ is the vector dimension. $k_m^{t, [\text{CLS}]}$ 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 $c_m^t \in \mathbb{R}^d$ and knowledge $k_m^t \in \mathbb{R}^d$, by the averaging all the token hidden states in their positions, this process for each pair input $I_m$ is denoted as:

$$\bar{K}_m^t, k_m^{t, [\text{CLS}]}; c_m^t, k_m^t = \text{BERT}(I_m)$$ (2)

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

$$h_m^t = \tanh(W_c c_m^t)$$

$$\alpha_m^t = \frac{\exp(V_c h_m^t)}{\sum_{i=1}^M \exp(V_c h_i^t)}, \ c^t = \sum_{i=1}^M \alpha_i^t h_i^t$$ (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 $k_m^{t, [\text{CLS}]}$ 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 $v_{coh}^{t, m} \in \mathbb{R}^2$. This entailing feature $v_{coh}^{t, m}$ 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 $k_m^{t-1, [\text{CLS}]}$. Inspired by (Chen et al., 2017), we apply the cross operator $f(u, v) = [u - v; u \odot v]$, 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 \([C_t; d^c_t]\) 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 \(d^k_t\) by gumbel-softmax. The below gray part is the GPT-2 based decoder with context \(C_t\) and \(d^k_t\) as input.

last chosen knowledge \(k_{gt}^{t-1}\) and each candidate knowledge \(k^t_{m,[CLS]}\), denoted as:

\[
q^t_m = f(k_{gt}^{t-1}; k^t_{m,[CLS]})
\]

\[
v^{cro}_{t,m} = \text{tanh}(F(C(q^t_m))), \quad m \in [1, ..., M]
\]  

(4)

We set \(v^{cro}_{t,m}\) 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 \(v^{coh}_{t,m}\) sequential knowledge difference \(v^{cro}_{t,m}\), context-aware knowledge embedding \(k^t_{m,[CLS]}\) and knowledge embedding \(k^t_m\), as described above, together form the transition-aware knowledge representations \(E^t\) for \(M\) relevant knowledge, which is denoted as:

\[
e^t_m = [v^{coh}_{t,m}; v^{cro}_{t,m}; k^t_{m,[CLS]}; k^t_m]
\]

\[
E^t = (e^t_1, ..., e^t_M)
\]  

(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 \(k_{gt}^{t-1}\), \(k^t_g\) and \(k^t_S\) to compute the information variance of two tuples, \(\langle k_{gt}^{t-1}, k^t_g \rangle\) and \(\langle k_{gt}^{t-1}, k^t_S \rangle\). 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:

\[
L_{sc} = D_{KL}(F(k_{gt}^{t-1}, k^t_S) \parallel F(k_{gt}^{t-1}, k^t_g))
\]  

(6)

We define the variance measure function as:

\[
F(u, v) = \log_{\text{softmax}}(\langle [u - v]^2; u \otimes v \rangle)
\]  

(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 \(G\) is constructed based on the text-similarity (tf-idf) of the candidate knowledge.
sentences. Each node is initialized by the knowledge representation $k_m^l$ from the BERT encoder. The output embedding of all nodes is then fed into an average pooling layer to obtain the graph representation $h^t$, capturing the relationship and semantic features of all knowledge candidates related to the dialog context. The process is formulated as:

$$h^t = \text{avgpool}(\text{GAT}([k_m^l]_M, G))$$  

(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 $c^t$ as query, transition-aware knowledge representations $E^t$ as keys, and knowledge interactive representation $h^t$ as the encoder hidden state, the Ptr-Net decodes knowledge as:

$$\hat{h}^t = \text{LSTMCell}(c^t, h^t)$$

$$\hat{\alpha}_m^t = v^\top \text{tanh}(W_c e_m^t + W_h h^t + b)$$

$$P(d_m^t | U_i, D_i) = \frac{\exp(\hat{\alpha}_m^t)}{\sum_{i=1}^{M} \exp(\hat{\alpha}_i^t)}$$

(9)

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

In a knowledge-grounded conversation dataset, the turns of all the conversations form the samples $\mathcal{D} = \{(U_i, D_i, r_i)\}_{i=1}^{N}$ where $U_i, D_i = \{d_m^l\}_M$ 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:

$$L_{cls} = L_{ce} + \lambda L_{sc}$$

$$L_{ce} = - \frac{1}{N} \sum_{i=1}^{N} \hat{y}_l \log(p(d_l^i | U_i, D_i))$$

(10)

where $\hat{y}_l$ denotes the one-hot vector indicating the ground-truth knowledge for data sample $i$, $p(d_l^i | U_i, D_i)$ denotes the probability distribution over the candidate knowledge $D_i$, $L_{ce}$ is a standard cross-entropy loss function for knowledge selection and $\lambda$ 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 $d_l^i$ 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:

$$p(r_i^j | U_i, d_l^j, r_{<i}^j) = \text{GPT-2}(d_l^j, U_i, r_{<i}^j)$$

$$L_g = - \frac{1}{N} \sum_{i=1}^{N} \sum_{|r_i|} \hat{y}_r^i \log p(r_r^i | U_i, d_l^r, r_{<r}^i)$$

(11)

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

### 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.
We compare our C2T2 with most of the existing 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 DI-ALKI (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 DI-ALKI, reaching 34.5/35.6 in terms of ACC on Test Seen and Unseen of WoW.

### 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.

### 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 Transformers¹ (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 $\lambda$ and $\mu$ for $L_{sc}$ and $L_g$ to 0.5 and 2. We will release all the codes and hyper-parameters setting for re-production.

### 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

¹https://github.com/huggingface/transformers
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 CoLVi 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 KnowledGPT 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 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 $c_t$ and each transition-aware knowledge representations $E_k$. 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.
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 knowledgedGPT, although C2T2 and knowledgedGPT both used GPT-2 as the generator, C2T2 still performed better. We also computed the FLewis’s 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, knowledgedGPT 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, knowledgedGPT 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.
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


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


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
We also calculate ROUGE scores on WoW dataset. 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.

### 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 2nd to 5th 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 4th to 5th 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 dataset. 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 seemingly 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 movies. 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.
<table>
<thead>
<tr>
<th>Topic: Halloween.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Post 1:</strong> Halloween is the best holiday ever!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth response</th>
<th>C2T2</th>
<th>KnowledGPT</th>
<th>SLKS</th>
<th>DukeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>(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.</td>
<td>(k1) I love it! It’s a celebration observed in a number of countries on 31 October, the eve of the western Christian feast.</td>
<td>(k1) I love Halloween, also known as allhallows, I love it too!</td>
<td>(k1) I love Halloween! I love that it is an annual festival!</td>
<td>(k1) I love Halloween! It’s a great place to visit.</td>
</tr>
</tbody>
</table>

| **Post 2:** Yes, I love being able to dress up and watch scary movies. |

<table>
<thead>
<tr>
<th>Ground-truth response</th>
<th>C2T2</th>
<th>KnowledGPT</th>
<th>SLKS</th>
<th>DukeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k2) It begins the three-day observance of allhallow tide, the time in the liturgical year dedicated to remembering the dead, including saints.</td>
<td>(k2) Halloween is the 3 day observance of allhallowtide, the time in the liturgical year dedicated to remembering the dead, including saints</td>
<td>(k4) I love watching scary movies. Halloween traditions originated from ancient Celtic harvest festivals, particularly the Gaelic festival Samhain.</td>
<td>(k3) Yes, me too! Trick or treating is so cool!</td>
<td>(k3) I love Halloween too! I like to watch trick or treating pumpkins.</td>
</tr>
</tbody>
</table>

| **Post 3:** I never knew that! What culture is that celebrated in? |

<table>
<thead>
<tr>
<th>Ground-truth response</th>
<th>C2T2</th>
<th>KnowledGPT</th>
<th>SLKS</th>
<th>DukeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k4) It is widely believed that many Halloween traditions originated from ancient Celtic harvest festivals, particularly the Gaelic festival Samhain; that such festivals</td>
<td>(k4) It is widely believed that many Halloween traditions originated from ancient Celtic harvest festivals, particularly the Gaelic festival Samhain; that</td>
<td>(k5) It is celebrated in many parts of the world.</td>
<td>(k2) It is celebrated the three day observance of allhallowtide, the time in the liturgical year dedicated to remembering the dead (1).</td>
<td></td>
</tr>
</tbody>
</table>

| **Post 4:** That would make sense. What about the Halloween movies with Michael Myers? Do you like those? |

<table>
<thead>
<tr>
<th>Ground-truth response</th>
<th>C2T2</th>
<th>KnowledGPT</th>
<th>SLKS</th>
<th>DukeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k8) 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.</td>
<td>(k6) I love Halloween, it is a 1978 American slasher film directed and scored by John Carpenter, co-written with producer Debra Hill, and starring Donald Pleas</td>
<td>(k6) I love the Halloween movies with Michael Myers - it’s so much fun!</td>
<td>(k9) I love Michael Carpenter. He was a young boy. He was a young boy.</td>
<td>(k6) I love those movies. They’re my favorites to watch on Halloween.</td>
</tr>
</tbody>
</table>

| **Post 5:** I love horror films, they elicit a physiological reaction, such as an elevated heartbeat, through the use of fear and shock. |

<table>
<thead>
<tr>
<th>Ground-truth response</th>
<th>C2T2</th>
<th>KnowledGPT</th>
<th>SLKS</th>
<th>DukeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k7) I love horror films, they elicit a physiological reaction, such as an elevated heartbeat, through the use of fear and shock</td>
<td>(k6) is a 1978 American slasher film directed and scored by John Carpenter, co-written with producer Debra Hill, and starring Donald Pleas</td>
<td>(k6) It was directed by John Carpenter</td>
<td>(k6) I love the movie Halloween. It’s a great American slasher.</td>
<td></td>
</tr>
</tbody>
</table>

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.