Knowledge Graph Grounded Goal Planning for Open-Domain Conversation Generation

Jun Xu,[1] **Haifeng Wang,**[2] **Zhengyu Niu,**[2] **Hua Wu,**[2] **Wanxiang Che**[1]*
[1]Harbin Institute of Technology, Harbin, China
[2]Baidu Inc., Beijing, China
{jxu, car}@ir.hit.edu.cn, {wanghaifeng, niuzhengyu, wu_hua}@baidu.com

Abstract

Previous neural models on open-domain conversation generation have no effective mechanisms to manage chatting topics, and tend to produce less coherent dialogs. Inspired by the strategies in human-human dialogs, we divide the task of multi-turn open-domain conversation generation into two sub-tasks: explicit goal (chatting about a topic) sequence planning and goal completion by topic elaboration. To this end, we propose a three-layer Knowledge aware Hierarchical Reinforcement Learning based Model (**KnowHRL**). Specifically, for the first sub-task, the upper-layer policy learns to traverse a knowledge graph (KG) in order to plan a high-level goal sequence towards a good balance between dialog coherence and topic consistency with user interests. For the second sub-task, the middle-layer policy and the lower-layer one work together to produce an in-depth multi-turn conversation about a single topic with a goal-driven generation mechanism. The capability of goal-sequence planning enables chatbots to conduct proactive open-domain conversations towards recommended topics, which has many practical applications. Experiments demonstrate that our model outperforms state of the art baselines in terms of user-interest consistency, dialog coherence, and knowledge accuracy.

Introduction

As letting machines talk with humans is one of the goals of AI, lots of research efforts have been devoted to open-domain conversation generation (Shang, Lu, and Li 2015). However, these Seq2Seq based models tend to produce generic or less coherent responses. To address this issue, previous studies introduce external knowledge (Liu et al. 2018; Zhou et al. 2018) or topic information (Wang et al. 2018; Xing et al. 2017) to improve dialog informativeness. Moreover, there are other studies to employ reinforcement learning (RL) with the aim of generating coherent and long-lasting multi-turn dialogs (Li et al. 2016b; Yao et al. 2018; Zhang et al. 2018; Zhao, Xie, and Eskenazi 2019).

Although these models have achieved promising results, they still tend to produce less coherent dialogs with loosely-connected topics especially for a long conversation. It may be explained by that they have no modules or effective mechanisms to manage chatting topics to ensure dialog coherence. Previous study (Hirano and Matsuo 2016) indicates that dialog management strategies are quite universal in human-human dialogs, e.g., topic changing or topic elaboration. Therefore, chatting topic management is crucial to generating coherent dialogs. Moreover, it is quite challenging to learn the decision-making process for topic management merely from dialog data without the help of background knowledge.

To address these issues, there are **two key challenges**. **The first one** is how to conduct high-level goal[1] sequence planning. This is difficult in that the chatbot should maintain inter-topic coherence[2], and at the same time it should also take consideration of user interests for goal decision to avoid "one-sided" conversation. **The second one** is how to generate an in-depth multi-turn conversation about a single topic for goal completion, which corresponds to intra-topic coherence[2]. The capability of goal-sequence planning enables chatbots to conduct proactive open-domain conversations towards recommended topics (Moon et al. 2019; Li et al. 2018; Tang et al. 2019; Wu et al. 2019).

To this end, we present a three-layer Knowledge graph grounded Hierarchical RL model (**KnowHRL**) for **goal-sequence planning** and **goal completion** by topic elaboration. The upper-layer policy corresponds to the first challenge, and the other two policies correspond to the second challenge. The **upper-layer** policy is a Multilayer Perceptron (MLP) based model, which learns natural paths over a knowledge graph (KG) in order to plan a coherent, diverse and long-lasting goal sequence. At the same time, we

*Corresponding author

[1]In this paper, goal refers to in-depth conversation about a given topic, which is an entity in a knowledge graph. It should be noted that the propose framework can be extended to cover more kinds of goals, e.g., recommending items by conversation, or booking a ticket, and so on, if we augment the KG to incorporate related resource for these types of goals.

[2]In this work, inter-topic coherence indicates that the whole conversation is composed of well-connected topics as a sequence, and intra-topic coherence indicates that each topic is elaborated by well-connected consecutive utterances.

Goals	Topic Facets	Utterances
<Between Calmness and Enthusiasm>	Start_of_Session	M: Do you like love movies? U: Yes, I often watch love movies recently. M: Then you can go to see the movie <Between Calmness and Enthusiasm>(2001). It's a love movie. U: I've seen this movie. There are many cheaters in it. Another movie with the same rating is <The World Is Not Enough>(1999).
<The World Is Not Enough>	Starring, Robert Carlyle Starring, Pierce Brosnan	M: Robert Carlyle starred in this movie, whose constellation is Aries. U: Yes, he starred. His ancestral home is Glasgow, Scotland. M: The film is starred by Pierce Brosnan, who was born in May 1953. Do you know who it is? U: Of course. He starred with Robert Carlyle and won the 24th Saturn Award for Best Actor.

Table 1: A dialog example between KnowHRL and a human, in which there is a goal sequence "<Between Calmness and Enthusiasm>, <The World Is Not Enough>". "M" is KnowHRL and "U" is a human. We see that KnowHRL can plan a goal sequence to support chatting topic management, and at the same time it can also respond appropriately to the human especially when the human mentions a new topic, e.g., "<The World Is Not Enough>" in the example.

consider user feedback information (e.g., new topics[3] from users) in reward functions to reinforce goal sequences being consistent with user's interests. The **middle-layer** policy is another MLP based model, which learns to select optimal neighboring vertices around the goal vertex as topic facets for an in-depth multi-turn conversation. Then the **lower-layer** policy uses a multi-mapping based neural generation model (Chen et al. 2019) to produce a multi-turn dialog conditioned on user utterances and topic facets. Thus we provide a goal-driven generation mechanism that is composed of the topic-facet selection operation and the use of topic-facets to guide generation, which can guarantee the completion of the second sub-task. Finally, we employ two models, the KnowHRL and a user simulator, to explore the space of possible actions. The KnowHRL is trained by optimizing long-term developer-defined rewards with advantage actor-critic method (A2C) (Sutton and Barto 2018). Table 1 provides a dialog example between KnowHRL and a human.

Evaluations against both user simulator and human subjects demonstrate the effectiveness of KnowHRL in terms of user-interest consistency, dialog coherence, and knowledge accuracy, when compared with state-of-the-art baselines.

Our contribution is summarized as follows:

- This work is the first attempt to divide the task of multi-turn open-domain conversation generation into two sub-tasks: goal-sequence planning, and goal completion by topic elaboration. Following this strategy, we propose the KnowHRL model.
- With the help of KG, we introduce explicit explainable dialog states and actions for policy learning. It brings the two benefits: (1) it is convenient to design goal related rewards to optimize the planning of goals and facets, (2) we use the information of goals and facets to guide response generation for better coherence and informativeness.
- Experiments demonstrate the effectiveness of KnowHRL in terms of user-interest consistency, dialog coherence, and knowledge accuracy.

[3]The topics that have no clear connection with current topic.

The Proposed Model

Problem Definition

Towards the aim of conducting a coherent multi-turn human-machine dialog, we divide the task of open-domain conversation generation into two sub-tasks: (1) goal-sequence planning, and (2) goal completion by generating an in-depth conversation about a topic. We formulate such a hierarchical decision making process within the Options framework (Sutton, Precup, and Singh 1999), which is closely related to Semi-Markov Decision Process (SMDP). With options, the agent can choose a "multi-step" action rather than only choose a primitive action at each time step in the traditional MDP settings. We treat goals as options where goals naturally require multiple steps to be accomplished. Next we will provide more details.

The KnowHRL Model

The overview of KnowHRL is shown in Figure 1. It has three hierarchical policies, which can address two sub-tasks: goal-sequence planning and goal completion by topic elaboration.

For the first sub-task, the upper-layer policy learns to traverse a KG in order to plan a goal sequence. It is difficult in that there should be a good balance between dialog coherence and user-interest consistency. Given a vertex from the KG as current chatting topic, the upper-layer policy learns to select an optimal chatting topic from all the one-hop neighbors of current vertex and new topics mentioned by the user.

For the second sub-task, it is quite challenging to conduct an in-depth conversation about a given chatting topic, especially for low-resource topics. Therefore we conduct in-depth conversation in two steps. Firstly, the middle-layer policy selects one of one-hop neighbors of current goal vertex as a topic facet. Then we use both the given goal and one of its topic-facets to guide the lower-layer policy in order to generate a multi-turn in-depth conversation.

Model Overview As shown in Figure 1, at time step $(t, 0, 0)$ (shorten as (t)), the upper-layer policy μ^{up} obtains the state s_t from the environment and selects a goal g_t for

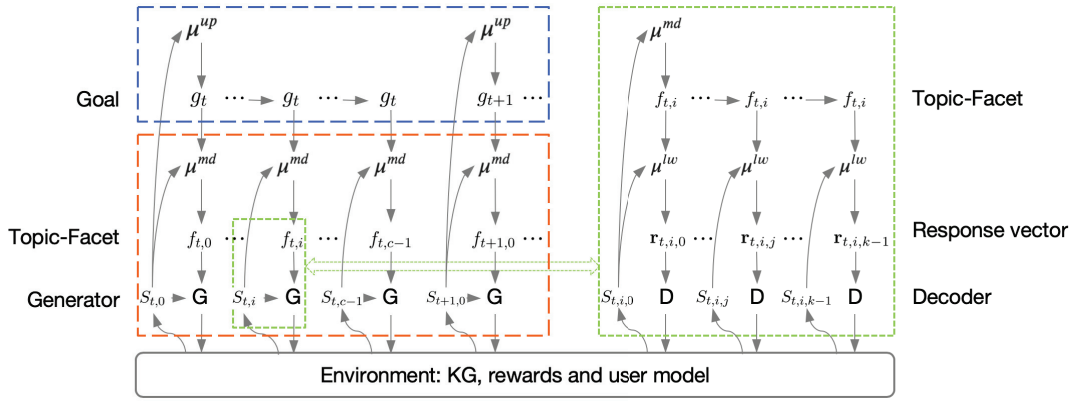

Figure 1: Overview of our model-KnowHRL. Goal-sequence planning is conducted by the upper-layer policy μ^{up} as shown in the blue-dotted rectangle. In-depth conversation about a topic as shown in the orange-dotted rectangle is conducted with two steps: (1) the middle-layer policy μ^{md} selects facets about a topic, and (2) the lower-layer policy μ^{lw} generates a multi-turn conversation about a facet $f_{t,i}$, with more details as shown in the green-dotted rectangle. Meantime, $S_t = S_{t,0}$ and $S_{t,i} = S_{t,i,0}$.

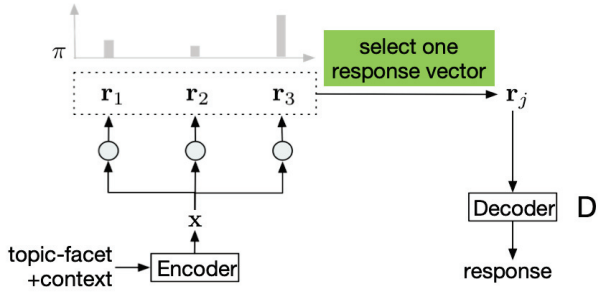

Figure 2: The multi-mapping based generation model. After the upper-layer and middle-layer policies make decisions to select goal and topic-facet, at each time step, the facet and corresponding context utterances are concatenated to calculate response vectors $\mathbf{r}$. The lower-layer policy selects one response vector which is further decoded into a response.

the middle policy μ^{md} and the lower policy μ^{lw}. The goal g_t is produced by sampling from its policy $g_t \sim \mu^{up}$ when the previous goal is finished.

Towards the goal g_t, the middle layer policy and the lower-layer policy work together to produce an in-depth conversation about a set of topic-facets. At time step $(t, i, 0)$ (shorten as (t, i)), the middle layer policy μ^{md} observes the state s_t and the goal g_t, and then selects a facet $f_{t,i}$ about g_t to start (or continue) an in-depth conversation. For example, in a conversation about *X-Men*, it is reasonable to talk about "who is its director" (*Bryan Singer* as the answer). In this case, the facet is a knowledge triple "[*X-Men*, Director, *Bryan Singer*]". With $f_{t,i}$ as the topic facet, at time step (t, i, j), the lower layer policy μ^{lw} selects a response vector $v_{t,i,j}$ for utterance generation through a pre-trained decoder (a multi-mapping based generator (Chen et al. 2019)). Then with an utterance from KnowHRL as the input, the user simulator will generate an appropriate response. This dialog about $f_{t,i}$ between KnowHRL and the user simulator will continue till the attribute name from the triple of $f_{t,i}$ is mentioned or the conversation reaches the maximum number of turns. The utterances generated from time step $(t, 0)$ to $(t, c-1)$ constitute an in-depth dialog for g_t. Thus we provide a goal-driven generation mechanism to guarantee the completion of topic-elaboration task, which consists of topic facet selection operation and the use of topic facets to guide generation.

State The state S consists of goal history g, topic-facet history f, context utterances u and a special symbol ut indicating whether the user mentions a new topic in the last utterance (ut = 1) or not (ut = 0). In this work, we choose the previous two utterances as u.

At the upper-layer, the chosen goal (topic) should be closely related to current goal (topic) to ensure inter-topic coherence of a dialog. Therefore, we use all the neighboring vertices of current goal vertex to constitute an action space. Moreover, to personalize the chatting topics for the user, we also include all the topics mentioned by the user.

At the middle-layer, its action space consists of all the neighboring vertices of current goal vertex except the vertices that have been talked about before.

At the lower-layer, its action space is infinite as arbitrary-length utterance can be chosen, which brings difficulty to policy learning (Li et al. 2016b). To filter the action space, we use a set of response vectors as actions, each of which represents a typical way to respond a given context, e.g. responding with interrogative sentences. Specifically, with concatenated topic-facet and contextual utterances encoded into $\mathbf{x}$ as input, multiple MLP-based networks will transform it into a set of response vectors $\{\mathbf{r}_j\}_{j=1}^{NL_r}$, where NL_r is the number of mapping networks. Each $\mathbf{r}_j$ can be further decoded into a response. This process is shown in the left part of Figure 2.

Policy For the upper-layer policy, at time step $(t, 0, 0)$ (shorten as (t)), we utilize three RNN encoders to represent

state S as $\mathbf{s}_t^{up} = \mathbf{W}^{up}[\mathbf{s}_{u,t}^{up}; \mathbf{s}_{g,t}^{up}; \mathbf{s}_{f,t}^{up}; ut]$, and μ^{up} is defined by:

$$\mu^{up}(\mathbf{s}_t^{up}, \mathbf{v}_{g_m}) = \frac{\exp((\mathbf{s}_t^{up})^T \mathbf{v}_{g_m})}{\sum_{l=1}^{NT_g} \exp((\mathbf{s}_t^{up})^T \mathbf{v}_{g_l})}, \quad (1)$$

where $\mathbf{W}^{up}$ is a weighting matrix, $\mathbf{v}_{g_m} = [\mathbf{e}_{g_m}; nu]$ is the concatenation of embedding of the m-th goal candidate $\mathbf{e}_{g_m}$ and nu that indicates whether g_m is the new topic mentioned by the user (nu = 1) or not (nu = 0), and NT_g is the number of goal candidates.

For the middle-layer policy, at time step (t, i), we utilize another three RNN encoders to represent the state as $\mathbf{s}_{t,i}^{md} = [\mathbf{s}_{u,t,i}^{md}; \mathbf{s}_{g,t,i}^{md}; \mathbf{s}_{f,t,i}^{md}]$, and the policy μ^{md} is defined by:

$$\mu^{md}(\mathbf{s}_{t,i}^{md}, \mathbf{e}_{f_{t,m}}) = \frac{\exp((\mathbf{s}_{t,i}^{md})^T \mathbf{e}_{f_{t,m}})}{\sum_{l=1}^{NT_f} \exp((\mathbf{s}_{t,i}^{md})^T \mathbf{e}_{f_{t,l}})}, \quad (2)$$

where $\mathbf{e}_{f_{t,m}}$ stands for the embedding of the m-th facet candidate (by concatenating embeddings of three parts in a knowledge triple), and NT_f is the number of topic-facet candidates. Meantime, as the completion of given goal is formulated as "option" in SMDP, we compute a flag to indicate whether we have finished current goal ($flag$ = 1) or not ($flag$ = 0) based on current state. Concretely, Three RNN encoders with LSTM units are utilized to represent state S as $\mathbf{s}_{t,i} = [\mathbf{s}_{u,t,i}; \mathbf{s}_{g,t,i}; \mathbf{s}_{f,t,i}; ut]$, and we decide the completion of the given goal with probability distribution calculated by:

$$p(flag|\mathbf{s}_{t,i}) = sigmoid(\mathbf{W}_a^T \mathbf{s}_{t,i}), \quad (3)$$

where $\mathbf{W}_a$ is a weighting matrix. Together, We define the probability of deciding action $f_{t,i}$ at step (t, i) as: $p(f_{t,i})$ = $p^{md}(f_{t,i}) * p(flag_{t,i} = 0)$.

For the lower-layer policy, at time step (t, i, j), another three RNN encoders are utilized to represent the state as $\mathbf{s}_{\mathbf{t,i,j}}^{\mathbf{low}} = [\mathbf{s}_{u,t,i,j}^{lw}; \mathbf{s}_{g,t,i,j}^{lw}; \mathbf{s}_{f,t,i}^{md}]$, and policy μ^{lw} is defined by:

$$\mu^{lw}(\mathbf{s}_{t,i,j}^{lw}, \mathbf{r}_m) = \frac{\exp((\mathbf{s}_{t,i,j}^{lw})^T \mathbf{r}_m)}{\sum_{l=1}^{NL_r} \exp((\mathbf{s}_{t,i,j}^{lw})^T \mathbf{r}_l)}, \quad (4)$$

where $\mathbf{r}_m$ stands for the m-th response vector candidate, and NL_r is the total number of response vectors.

Multi-mapping generator To capture typical ways for a model to respond, we employ a multi-mapping based generator proposed by (Chen et al. 2019), shown in Figure 2.

First, the topic-facet triple[4] and contextual utterances are concatenated and then fed into a RNN context encoder for computation of the context vector $\mathbf{x}$. Then the generator maps $\mathbf{x}$ to candidate response representations $\{\mathbf{r}_j\}_{j=1}^{NL_r}$ through NL_r different mapping functions modeled by MLP networks[5]. Finally, only one candidate representation $\mathbf{r}_j$ will be selected and fed into a RNN decoder as initial hidden state for response generation. During training procedure, we utilize gumbel-softmax to sample from selection probability distribution π which is defined based on semantic similarity between target response and candidates. Formally,

$$\pi = \frac{\exp(\mathbf{r}_j^T \mathbf{y})}{\sum_{i=1}^{NL_r} \exp(\mathbf{r}_i^T \mathbf{y})}, \quad (5)$$

where $\mathbf{y}$ is obtained by encoding ground-truth response with a RNN response-encoder. Intuitively, this can be seen as a "clustering" process where targets are clustered into NL_r classes. Vector $\mathbf{r}_j$ from each cluster represents a typical response. After optimization, each vector $\mathbf{r}_j$ is assumed to be able to generate a high-quality response.

We introduce an auxiliary matching loss L_M to train the response-encoder. Particularly, for the encoded context vector $\mathbf{x}$ and the target response vector $\mathbf{y}$, another negative sample vector $\mathbf{y}^-$ is calculated by encoding a randomly sampled utterance from training set with the target encoder. L_M is defined as:

$$L_M = -log\sigma(\mathbf{x}, \mathbf{y}) + log\sigma(\mathbf{x}, \mathbf{y}^-), \quad (6)$$

where σ is a sigmoid function, and x and y are compared by dot product. The loss function of the multi-mapping based generator is defined as $L = L_G + L_M$, where L_G is the standard negative log-likelihood loss.

Rewards Next we provide the details of reward factors, denoted as r_{up}, r_{md} and r_{lw}, used for the three layers respectively.

For the upper-layer policy, we define its rewards r_{up} as a weighted sum of the following factors with weights $\{\alpha\}_1^5$.

- Coherence of the goal sequence. We calculate the average cosine distance between the chosen goal and one of history goals in graph embedding space (TransE (Bordes et al. 2013) trained on KG provided) as coherence reward.
- User interest consistency. This reward is 0 if the chosen goal is not identical with the new topic mentioned by the user, otherwise 1.
- Diversity. We should have a good balance between changing the goal too frequently and always sticking to the same goal. This reward is defined as 0 when the number of facets talked around given goal lies in interval [2,4], otherwise -1.
- Sustainability. It is reasonable to give priority to vertices with lots of related knowledge or topic facets, we represent this reward as PageRank score of the chosen goal vertex, which is calculated on the KG.
- Goal-completion information from the middle layer. We calculate as the average rewards of r_{md}.

For the middle layer, we define its rewards r_{md} as a weighted sum of the following factors with weights $\{\beta\}_1^2$.

- Topic-facet coherence. We calculate coherence as cosine distance in embedding space between the selected facet and current topic.
- Rewards from the lower layer. We define as the average rewards of r_{low}.

For the lower layer, we define its rewards r_{lw} as a weighted sum of the following factors with weights $\{\phi\}_1^3$.

[4]For the user simulator, only contextual utterances are used for computation of the context vector $\mathbf{x}$.

[5]MLPs are two-layer fully connected perceptron, with hidden layer size as "512". And the number of MLPs NL_r = 10.

- Utterance relevance. It is calculated as $r_{rel}^{lw} = \sigma(\mathbf{b}_t \cdot \mathbf{c}_t)$, where $\mathbf{b}_t$ and $\mathbf{c}_t$ stand for response and context encoded by a response-encoder and a context encoder in the multi-mapping generator, σ is a sigmoid function.
- Utterance informativeness. It is 1 if generated response contains at least one piece of knowledge, otherwise 0.
- Topic-facet completion. It is 0 if the lower layer fails to lead a conversation toward the given topic facet, or 1 if the model succeeds but without mentioning the relations between goals and facets, or 2 if the model succeeds and at the same time mentions the relations.

In our experiment[6], $\{\alpha\}_1^5$ are set as 1, 5, 1, 5000, 0.5; β_1 equals to 1 and β_2 is 0.5; $\{\phi\}_1^3$ are set as 1, 1, 2.

Optimization We employ the A2C method (Sutton and Barto 2018) for model optimization rather than original policy gradient as used in previous works (Li et al. 2016b) to make the learning process be stable. Moreover, we only update the parameters of policies, and the pre-trained multi-mapping generator and the user simulator stay intact.

Experiments and Results

Dataset

We use a publicly available knowledge-driven dialog dataset, DuConv[7], for pretraining of the multi-mapping based generator, baselines and the user simulator. The dataset consists of 30k dialogs with 120k dialogue turns. We split it into training set (100k-turn), development set (10k-turn) and test set (10k-turn). It also provides a KG in the domain of movies and celebrities. Each dialog is annotated by two crowd-sourced annotators to conduct a multi-turn KG grounded conversation towards a given entity, where the two humans play the roles of "chatbot" and "user" respectively.

As the proposed KnowHRL have clear explainable dialog states, e.g., explicit goals and facets, it enables us to employ a "label trick" strategy to modify the train/development/test set to ensure knowledge accuracy in generated utterances. As shown in Figure 3: (1) we replace each topic word and each topic-facet word with the label "topic" and "facet" respectively. Meantime, we replace other triple values mentioned in the corpus with corresponding triple attributes or relations; (2) we train KnowHRL and the multi-mapping generator on this transformed DuConv dataset, not the original one; (3) during testing procedure, KnowHRL generate utterances with labels; (4) we restore the original values to get final response utterances. This strategy is not applicable to baselines since they cannot provide what is current topic at each dialog turn, and then it is impossible to retrieve the correct triple values for "knowledge labels" in generated responses. Thus baseline models and the user simulator are trained on the original DuConv dataset.

Models

For empirical comparison, we select two state-of-the-art models that are closely related to ours as baselines, which also employ KG or RL for conversation generation. We implement the two baselines by ourselves.

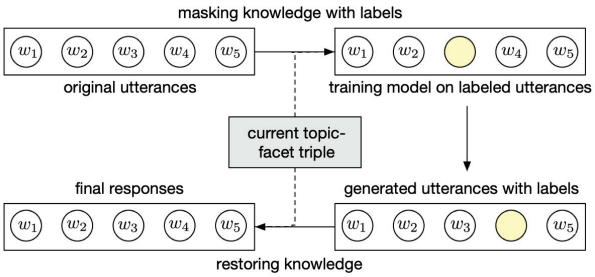

Figure 3: Data flow with the "label" trick for KnowHRL.

CCM It is a state-of-the-art KG based conversation model (Zhou et al. 2018), which can attentively read multiple knowledge triples to facilitate better response generation.

CCM+LaRL (Zhao, Xie, and Eskenazi 2019) proposed a latent variable based RL model (LaRL) for conversation generation, which outperforms traditional RL based dialog models (Li et al. 2016b) by a large margin. We choose the multivariate categorical latent variables as dialog actions which perform the best in their study. Furthermore, we enable LaRL to utilize KG by incorporating the static and dynamic graph attention mechanisms from CCM. This strong baseline is denoted as CCM+LaRL. We utilize the reward functions that use only utterance information for CCM+LaRL. We utilize relevance of utterances and informativeness for RL training.

KnowHRL It is the proposed model. We pretrain a multi-mapping generator with DuConv by choosing "bot" utterances as responses to serve as the foundation of KnowHRL.

KnowHRL-liteReward To verify the effectiveness of goal related reward factors for policy learning, we conduct an ablation study by our model without goal related factors. Concretely, we only use utterance relevance and utterance informativeness as rewards for RL training.

The **user simulator** is another multi-mapping based generation model trained on DuConv to predict user-side utterances. We use the same user simulator for RL training of the three models, CCM+LaRL, KnowHRL and KnowHRL-liteReward. During testing procedure, all the models share the same pre-trained user simulator.

Evaluation Settings

Conversation with User Simulator Following the experimental settings in prior work (Li et al. 2016b), we use the user simulator to play the role of human. Then we let each of the models to be evaluated generate the first utterance and chat with the user simulator, till they reach the maximum number of turns (which is set as 7 turns in this work). To check the capability of these models to consider user feedback, we randomly insert utterances with new entities (far away from current topic in KG) as user's responses, once for each session. Finally we collect multi-turn dialogs generated by each model for evaluation.

[6]We conduct grid search for weights estimation.

[7]More details at https://arxiv.org/abs/1906.05572

Conversation with Human We also perform evaluation against human subject for a more thorough empirical study. Specifically, we setup human evaluation interfaces and ask human annotators to converse with each of the models till they reach the maximum number (set as 10) of turns. To check the capability of these models to consider user feedback, we ask human to change chatting topics in the middle of session (e.g., at the third or fourth turn). Finally we collect multi-turn dialogs generated by each model for evaluation.

Evaluation Metrics

Metrics such as BLEU and perplexity have been widely used for dialog quality evaluation (Li et al. 2016a; Serban et al. 2016), but it is widely debated how well these automatic metrics are correlated with true response quality(Liu et al. 2016). Since the proposed model does not aim at predicting the highest-probability response at each turn, but rather the long-term success of a dialog, we do not employ BLEU or perplexity for evaluation, and we propose the following evaluation metrics.

Evaluation Metrics at Session Level For each model to be evaluated, we randomly sample 100 dialogs for the settings of conversation with user simulator or human, in which, on average, each dialog consists of 14 utterances. We ask three annotators to judge session-level quality. For the convenience of evaluation, each dialog is split into subsessions according to chatting topics, and then they judge the quality of each sub-session. Notice that model identifiers are masked during evaluation. Session-level metrics include:

- Intra-topic coherence. We first define intra-topic incoherence problems as follows: (1) anaphora errors across utterances, e.g. using "she" to refer to a male actor mentioned, (2) simply copying consecutive words or phrases within an utterance, (3) incorrect collocation, e.g., using another movie name as the director of the movie mentioned in previous utterances. For each sub-session, the annotators are asked to rate with a score of {"0", "+1", "+2"}. A sub-session will be rated "0" if it contains more than one incoherence problems, and then this sub-session is considered to be incomprehensible. If a sub-session contains one incoherence flaws, it will be rated "+1". A sub-session with no incoherence flaws will be rated "+2". Finally we obtain an average score for each dialog.
- Inter-topic coherence. For evaluation of topic-changing smoothness, annotators perform the judgment at each topic-changing position between two sub-sessions. They use intra-topic coherence results to help this annotation. Specifically, if the previous adjacent sub-session is incomprehensible ("0" for intra-topic coherence evaluation), then current position will be rated "0". Otherwise, current position will be rated "+2" if a new topic is introduced by mentioning its relation to current topic or topic facet. If only conjunction words (e.g. another) are used for topic changing, it will be rated "+1". It will be rated with "0" if a new topic suddenly appears without any sign or suggestion, and breaks current conversation flow. Finally we obtain an average score for each dialog.

Model	Intra.	Inter.	Dist-2	K.A.	Cons.
CCM	0.67	0.32	0.27	0.14	0.06
CCM+LaRL	0.94	0.44	0.32	0.17	0.09
KnowHRL	**1.42**	**1.39**	**0.39**	**0.88**	**0.89**
-liteReward	1.13	0.41	0.35	0.81	0.07

Table 2: Results of session-level evaluations on dialogs with user simulator. KnowHRL outperforms all the baselines significantly (sign test, p-value < 0.01) in all the metrics.

Model	Intra.	Inter.	Dist-2	K.A.	Cons.
CCM	0.72	0.40	0.29	0.13	0.09
CCM+LaRL	0.98	0.52	0.33	0.19	0.17
KnowHRL	**1.40**	**1.45**	**0.41**	**0.90**	**0.93**
-liteReward	1.16	0.46	0.38	0.84	0.13

Table 3: Results of session-level evaluations on dialogs with human. KnowHRL outperforms all the baselines significantly (sign test, p-value < 0.01) in terms of all the metrics.

- Distinct. The metric Dist-i calculates the ratio of distinct i-gram in generated responses (Li et al. 2016a). We use Dist-2 to measure the diversity of generated responses.
- Knowledge accuracy (K.A.). For each model, we randomly sample 100 entity mentions in generated utterances. Then the annotators judge their correctness based on triple attributes mentioned in utterances with the help of a KG and provide a score of {"0", "+1"}, where "1" means "correct knowledge in responses".
- User-interest consistency (Cons.). To evaluate if a model can respond appropriately when the user mentions a new topic, for each dialog, we ask the annotators to identify the positions with new topics and then judge the quality. A position will be rated "1" if the model follows the user's new topic and chats about it, otherwise "0".

Results As shown in Table 2 and Table 3, KnowHRL outperforms all the baselines significantly (sign test, p-value $<$ 0.01) in terms of all the metrics.

In terms of intra-topic coherence, KnowHRL outperforms baselines. KnowHRL can conduct an in-depth conversation with a clear "central topic", indicating the effectiveness of our goal-driven generation mechanism. However, the baselines tend to briefly talk about loosely related entities with lots of anaphora errors and incorrect collocations.

In terms of inter-topic coherence, KnowHR outperforms baselines by a large margin. KnowHRL can attain smooth dialog transition when changing the topics, indicating the effectiveness of our goal-sequence planning model and response generation model. However, CCM and CCM+LaRL tend to mention a new entity without indication of its relation to current topic. Moreover, they might use knowledge across multiple topics for generation at a single turn. In contrast, our model explicitly selects a topic and focuses on only one topic across multiple dialog turns.

In terms of Distinct-2, results show that our two model can generate responses with diverse knowledge triples. In

contrast, CCM and CCM+LaRL prefer to mention high frequency knowledge.

In terms of knowledge accuracy, results show that there is an issue of knowledge usage for the two baselines. The reason is that CCM selects knowledge based on similarity with various "attention" mechanisms, but there is no mechanism to guarantee the correctness of knowledge in generated responses. In contrast, response generation in KnowHRL is guided by goals and topic-facets, and our "label trick" strategy provides an promising mechanism to ensure correctness.

In terms of user-interest consistency, our models can generate responses being consistent with user's topics. In contrast, the two baselines tend to ignore new topics introduced by the user. It validates the effectiveness of our goal planning module. It also indicates that our model can achieve a good balance between global dialog coherence and local topic consistency with user interests.

Ablation study To clarify what boosts the performance of KnowHRL, we remove all the rewards from KnowHRL except the widely used two factors: utterance relevance and utterance informativeness, denoted as KnowHRL-liteReward. As shown in Table 2, the performances of KnowHRl-liteReward drops dramatically in terms of inter-topic coherence and user interest consistency. It indicates that goal related rewards are crucial to maintain inter-topic coherence and user interest consistency. Meantime, knowledge accuracy of KnowHRL-liteReward drops sightly, but still outperforms the two baseline models. It indicates that our advantage in knowledge accuracy is not brought by the rewards.

In summary, we have some findings: (1) With the help of KG, and our strategy of divide-and-conquer for open-domain conversation generation, we obtain explainable dialog states and actions for KnowHRL. (2) With these explainable states and actions, it is convenient to design goal-related rewards to optimize the planning of goals and topic facets, (3) These goals and topic facets can be used to guide response generation for better dialog coherence and informativeness. (4) With these explainable states and actions, KnowHRL is compatible with the "label trick" strategy, which leads to higher knowledge accuracy.

Evaluation at Turn Level For evaluation of each model, we randomly sample 200 utterances from dialogs with the setting of conversation with user simulator or human, and use the previous utterance as dialog context. We ask three annotators to judge turn-level quality. Notice that model identifiers are masked during evaluation. (1) Appropriateness. A response will be rated "0" if it is irrelevant to the context, otherwise "1". (2) Informativeness. A response will be rated "+1" if it contains at least a piece of knowledge from KG, e.g., entities/comments/facts, otherwise "0". As shown in Table 4, KnowHRL outperforms all the baselines in terms of turn-level response appropriateness and informativeness when chatting with user simulator and human.

Models	**Simulator**		**Human**	
	Appr.	Infor.	Appr.	Info.
CCM	0.74	0.81	0.79	0.84
CCM+LaRL	0.77	0.78	0.81	0.81
KnowHRL	**0.87**	**0.91**	**0.89**	**0.94**
-liteReward	0.82	0.83	0.85	0.87

Table 4: Turn-level results on dialogs with user simulator and dialogs with human.

Related Work

Seq2Seq Based Dialog Models To address the issue of generic responses in seq2seq models, some studies have been conducted to improve response informativeness (Yao et al. 2017; Xing et al. 2017; Wang et al. 2018; Zhao, Zhao, and Eskenazi 2017). However, these models have no explicit high-level topics to guide multi-turn conversation generation, thus tending to generate less coherent dialogs. Recently, imposing goals on open-domain conversation generation models having attracted lots of research interests (Moon et al. 2019; Li et al. 2018; Tang et al. 2019; Wu et al. 2019) since it enables practical applications, e.g., recommendation of engaging entities. However, these models can just produce a dialog towards a single goal, instead of a goal sequence as done in this work.

Dialog Models With RL RL has been used to encourage coherent, informative, and long-lasting utterance sequences (Li et al. 2016b; Serban et al. 2017; Zhang et al. 2018; Yao et al. 2018; Zhao, Xie, and Eskenazi 2019). However, they still tend to generate less coherent dialogs since they have no explicit high-level topics to guide response generation. Hierarchical RL models have been studied for task oriented dialog (Peng et al. 2017; Budzianowski et al. 2017), while we focus on open-domain dialog in this work.

Knowledge Aware Conversation Generation Further, there are growing interests in leveraging knowledge for generation of appropriate and informative responses (Ghazvininejad et al. 2018; Sangdo Han and Lee 2015; Liu et al. 2018; Vougiouklis, Hare, and Simperl 2016; Young et al. 2018; Zhou et al. 2018).

Conclusion

In this work we propose a Knowledge graph grounded Hierarchical RL based conversational model (**KnowHRL**) to demonstrate how hierarchical goal planning over a KG can facilitate chatting topic management and further response generation. Results show that KnowHRL outperforms baselines in terms of dialog coherence, user interest consistency, and knowledge accuracy. In the future, we would like to investigate how to enrich the content of KG to cover more chatting topics from open-domain dialog corpora.

Acknowledgments

This work is supported by the National Key Research and Development Project of China (No. 2018AAA0101900) and the National Natural Science Foundation of China (NSFC) via grant 61976072.

References

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, 2787–2795.

Budzianowski, P.; Ultes, S.; Su, P.-H.; Mrkšić, N.; Wen, T.-H.; Casanueva, I.; Rojas-Barahona, L.; and Gašić, M. 2017. Sub-domain modelling for dialogue management with hierarchical reinforcement learning. *arXiv preprint arXiv:1706.06210*.

Chen, C.; Peng, J.; Wang, F.; Xu, J.; and Wu, H. 2019. Generating multiple diverse responses with multi-mapping and posterior mapping selection. *Proceedings of IJCAI*.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; tau Yih, W.; and Galley, M. 2018. A knowledge-grounded neural conversation model. In *Proceedings of AAAI 2018*, 5110–5117.

Hirano, T., and Matsuo, R. H. Y. 2016. Analyzing post-dialogue comments by speakers-how do humans personalize their utterances in dialogue?-. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 157.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, 110–119.

Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; and Gao, J. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1192–1202.

Li, R.; Kahou, S. E.; Schulz, H.; Michalski, V.; Charlin, L.; and Pal, C. 2018. Towards deep conversational recommendations. In *Proceedings of NeurIPS*, 9748–9758.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132.

Liu, S.; Chen, H.; Ren, Z.; Feng, Y.; Liu, Q.; and Yin, D. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of ACL*, 1489–1498.

Moon, S.; Shah, P.; Kumar, A.; and Subba, R. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of ACL*.

Peng, B.; Li, X.; Li, L.; Gao, J.; Celikyilmaz, A.; Lee, S.; and Wong, K.-F. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2231–2240.

Sangdo Han, Jeesoo Bang, S. R., and Lee, G. G. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of SIGDIAL*, 129–133.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI*, 3776–3784.

Serban, I. V.; Sankar, C.; Germain, M.; Zhang, S.; Lin, Z.; Subramanian, S.; Kim, T.; Pieper, M.; Chandar, S.; Ke, N. R.; et al. 2017. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In *Proceedings of ACL-IJCNLP*, volume 1, 1577–1586.

Sutton, R. S., and Barto, A. G. 2018. *Reinforcement learning: An introduction*.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112(1-2).

Tang, J.; Zhao, T.; Xiong, C.; Liang, X.; Xing, E. P.; and Hu, Z. 2019. Target-guided open-domain conversation. In *Proceedings of ACL*.

Vougiouklis, P.; Hare, J.; and Simperl, E. 2016. A neural network approach for knowledge-driven response generation. In *Proceedings of COLING 2016*, 3370–3380.

Wang, W.; Huang, M.; Xu, X.-S.; Shen, F.; and Nie, L. 2018. Chat more: Deepening and widening the chatting topic via a deep model. In *Proceedings of SIGIR*.

Wu, W.; Guo, Z.; Zhou, X.; Wu, H.; Zhang, X.; Lian, R.; and Wang, H. 2019. Proactive human-machine conversation with explicit conversation goals. In *Proceedings of ACL*.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic aware neural response generation. In *AAAI*, volume 17, 3351–3357.

Yao, L.; Zhang, Y.; Feng, Y.; Zhao, D.; and Yan, R. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of EMNLP*, 2190–2199.

Yao, L.; Xu, R.; Li, C.; Zhao, D.; and Yan, R. 2018. Chat more if you like: Dynamic cue words planning to flow longer conversations. *arXiv preprint arXiv:1811.07631*.

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of AAAI 2018*, 4970–4977.

Zhang, W.-N.; Li, L.; Cao, D.; and Liu, T. 2018. Exploring implicit feedback for open domain conversation generation. In *Proceedings of AAAI*, 547–554.

Zhao, T.; Xie, K.; and Eskenazi, M. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1208–1218.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 654–664.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of IJCAI-ECAI*.