Learning Neural Templates for Recommender Dialogue System

Zujie Liang[1*], **Huang Hu**[2*], **Can Xu**[2], **Jian Miao**[2], **Yingying He**[2],
Yining Chen[2], **Xiubo Geng**[2], **Fan Liang**[1], **Daxin Jiang**[2†]
[1]Sun Yat-sen University, Guangzhou, China
[2]Microsoft Corporation, Beijing, China
[1]{liangzj9@mail2.sysu.edu.cn, isslf@mail.sysu.edu.cn}
[2]{huahu,caxu,jianm,yingyhe,yinichen,xigeng,djiang}@microsoft.com

Abstract

Though recent end-to-end neural models have shown the promising progress on Conversational Recommender System (CRS), two key challenges still remain. First, the recommended items cannot be always incorporated into the generated replies precisely and appropriately. Second, only the items mentioned in the training corpus have a chance to be recommended in the conversation. To tackle these challenges, we introduce a novel framework called **NTRD** for recommender dialogue system that decouples the dialogue generation from the item recommendation. **NTRD** has two key components, *i.e.*, response template generator and item selector. The former adopts an encoder-decoder model to generate a response template with slot locations tied to target items, while the latter fills in slot locations with the proper items using a sufficient attention mechanism. Our approach combines the strengths of both classical slot filling approaches (that are generally controllable) and modern neural NLG approaches (that are generally more natural and accurate). Extensive experiments on the benchmark REDIAL show our NTRD significantly outperforms the previous state-of-the-art methods. Besides, our approach has the unique advantage to produce novel items that do not appear in the training set of dialogue corpus. The code is available at https://github.com/jokieleung/NTRD.

1 Introduction

Building an intelligent dialogue system that can freely converse with human, and fulfill complex tasks like movie recommendation, travel planning and etc, has been one of longest standing goals of natural language processing (NLP) and artificial intelligence (AI). Thanks to the breakthrough in deep learning, the progress on dialogue system has been greatly advanced and brought into a new frontier over the past few years. Nowadays, we are witnessing the booming of virtual assistants with conversational user interface like Microsoft Cortana, Apple Siri, Amazon Alexa and Google Assistant. The recent large-scale dialogue models such as DialoGPT (Zhang et al., 2020), Meena (Adiwardana et al., 2020) and Blender (Roller et al., 2021), demonstrate the impressive performance in practice. Besides, the social bots such as XiaoIce (Shum et al., 2018) and PersonaChat (Zhang et al., 2018a) also exhibit the great potential on the emotional companion to humans.

The conversational techniques shed a new light on the search and recommender system, as the users can seek information through interactive dialogues with the system. Traditional recommender systems often rely on matrix factorization methods (Koren et al., 2009; Rendle, 2010; Wang et al., 2015; He et al., 2017), and suffer from the cold-start problem (Schein et al., 2002; Lika et al., 2014) when no prior knowledge about users is available. On the other hand, existing recommendation models are trained on offline historical data and have the inherent limitation in capturing online user behaviors (Yisong, 2020). However, the user preference is dynamic and often change with time. For instance, a user who usually prefers science fiction movies but is in the mood for comedies, would likely get a failed recommendation.

In recent years, there is an emerging trend towards building the recommender dialogue system, *i.e.*, Conversational Recommendation System (CRS), which aims to recommend precise items to users through natural conversations. Existing works (Li et al., 2018; Chen et al., 2019; Zhou et al., 2020a; Ma et al., 2020) on this line usually consist of two major components, namely a recommender module and a dialogue module. The recommender module aims at retrieving a subset of items

[*] Equal contribution. Work performed when Zujie Liang was an intern at Microsoft STCA.
[†] Corresponding author: djiang@microsoft.com.

that meet the user's interest from the item pool by conversation history, while the dialogue module generates free-form natural responses to proactively seek user preference, chat with users, and provide the recommendation. To incorporate the recommended items into the responses, a switching network (Gulcehre et al., 2016) or copy mechanism (Gu et al., 2016) is utilized by these methods to control whether to generate an ordinal word or an item at each time step. Such integration strategies cannot always incorporate the recommended items into generated replies precisely and appropriately. Besides, current approaches do not consider the generalization ability of the model. Hence, only the items mentioned in the training corpus have a chance of being recommended in the conversation.

In this paper, we propose to learn **N**eural **T**emplates for **R**ecommender **D**ialogue system, *i.e.*, **NTRD**. NTRD is a neural approach that firstly generates a response "template" with slot locations explicitly tied to the recommended items. These slots are then filled in with proper items by an item selector, which fully fuses the information from dialogue context, generated template and candidate items via the sufficient multi-head self-attention layers. The entire architecture (response template generator and item selector) is trained in an end-to-end manner. Our approach combines the advantages of both classical slot filling approaches (that are generally controllable) and modern neural NLG approaches (that are generally more natural and accurate), which brings both naturally sounded responses and more flexible item recommendation.

Another unique advantage of our NTRD lies in its zero-shot capability that can adapt with a regularly updated recommender system. Once a slotted response template is generated by the template generator, different recommender systems could be plugged into the item selector easily to fill in the slots with proper items. Thus, NTRD can produce the diverse natural responses with the items recommended by different recommenders.

The contributions of this work are summarized as follows: (1) We present a novel framework called NTRD for recommender dialogue system, which decouples the response generation from the item recommendation via a two-stage strategy; (2) Our NTRD first generates a response template that contains a mix of contextual words and slot locations explicitly associated with target items, and then fills in the slots with precise items by an item selector using a sufficient attention mechanism; (3) Extensive experiments on standard dataset demonstrate our NTRD significantly outperforms previous state-of-the-art methods on both automatic metrics and human evaluation. Besides, NTRD also exhibits the promising generalization ability on novel items that do not exist in training corpus.

2 Related Work

In this section, we first introduce the related work on task-oriented dialogue system. Then we review the existing literature on Conversational Recommender Systems (CRS), which can be roughly divided into two categories, *i.e.*, attribute-centric CRS and open-ended CRS.

Task-oriented Dialogue System. From the methodology perspective, there are two lines of the research on the task-oriented dialogue system, *i.e.*, modular approaches (Young et al., 2013) and end-to-end approaches (Serban et al., 2016; Wen et al., 2017; Bordes et al., 2017; Zhao et al., 2017; Lei et al., 2018). Recent works like GLMP (Wu et al., 2019) and dynamic fusion network (Qin et al., 2020) make attempt to dynamically incorporate the external knowledge bases into the end-to-end framework. Wu et al. (2019) introduce a global-to-local memory pointer network to RNN-based encoder-decoder framework to incorporate external knowledge in dialogue generation. By contrast, our approach gets rid of pointer network paradigm and proposes a two-stage framework, which is modeled by the transformer-based architecture.

Attribute-centric CRS. The attribute-centric CRS conducts the recommendations by asking clarification questions about the user preferences on a constrained set of item attributes. This kind of systems gradually narrow down the hypothesis space to search the optimal items according to the collected user preferences. The various asking strategies have been extensively explored, such as memory network based approach (Zhang et al., 2018b), entropy-ranking based approach (Wu et al., 2018), generalized binary search based approaches (Zou and Kanoulas, 2019; Zou et al., 2020), reinforcement learning based approaches (Sun and Zhang, 2018; Hu et al., 2018; Chen et al., 2018; Lei et al., 2020a; Deng et al., 2021; Li et al., 2021), adversarial learning based approach (Ren et al., 2020) and graph based approaches (Xu et al., 2020; Lei et al., 2020b; Ren et al., 2021; Xu et al., 2021). Most of these works (Christakopoulou et al., 2018;

Zhang et al., 2018b; Deng et al., 2021) retrieve questions/answers from a template pool and fill the pre-defined slots with optimal attributes. Although this kind of systems are popular in the industry due to the easy implementation, they are still lack of the flexibility and the interactiveness, which leads to the undesirable user experience in practice.

Open-ended CRS. Recently, researchers begin to explore the more free-style item recommendation in the response generation, *i.e.*, open-ended CRS (Li et al., 2018; Chen et al., 2019; Liao et al., 2019; Kang et al., 2019; Zhou et al., 2020a; Ma et al., 2020; Chen et al., 2020; Liu et al., 2020; Hayati et al., 2020; Zhou et al., 2020b; Zhang et al., 2021). Generally, this kind of systems consist of two major components, namely a recommender component to recommend items and a dialogue component to generate natural responses. Li et al. (2018) make the first attempt on this direction. They release a benchmark dataset REDIAL that collects human conversations about movie recommendation between paired crowd-workers with different roles (*i.e.*, Seeker and Recommender). Further studies (Chen et al., 2019; Zhou et al., 2020a; Ma et al., 2020; Sarkar et al., 2020; Lu et al., 2021) leverage multiple external knowledge bases to enhance the performance of recommendation. Liu et al. (2020) propose a multi-goal driven conversation generation framework (MGCG) to proactively and naturally lead a conversation from a non-recommendation dialogue to a recommendation-oriented one. Recently, Zhou et al. (2021) release an open-source CRS toolkit, *i.e.*, CRSLab, to facilitate the research on this direction. However, the Pointer Network (Gulcehre et al., 2016) or Copy Mechanism (Gu et al., 2016) used in these approaches cannot be always accurately incorporated the recommended items into the generated replies. Moreover, only the items mentioned in training corpus have a chance of being recommended in conversations by existing approaches.

Our work lies in the research of open-ended CRS. While in our work, we propose to decouple dialogue generation from the item recommendation. Our approach first leverages Seq2Seq model (Sutskever et al., 2014) to generate the response template, and then fills the slots in template with proper items using sufficient multi-head self-attention mechanism. Moreover, our work shows the unique advantage to produce novel items that do not exist in the training corpus.

3 Preliminary

Formally, a dialogue consisting of t-turn conversation utterances is denoted as $\mathcal{D} = \{s_t\}_{t=1}^{N}$. Let m denotes an item from the total item set $\mathcal{M}$, and w denotes a word from vocabulary $\mathcal{V}$. At the t-th turn, the recommender module chooses several candidate items $\mathcal{M}_t$ from the item set $\mathcal{M}$, while the dialogue module generates a natural language sentence s_t containing a proper item i from $\mathcal{M}_t$ to make recommendations. It is noteworthy that $\mathcal{M}_t$ can be equal to $\emptyset$ when there is no need for recommendation. In that case, the dialogue module could continue to generate a chit-chat response or proactively explore the user's interests by asking questions. To incorporate the recommended items into the generated reply, a switching mechanism (Gulcehre et al., 2016) or CopyNet (Gu et al., 2016) is usually utilized to control the decoder to decide whether it should generate a word from the vocabulary or an item from the recommender output. Specifically, the recommender predicts the probability distribution over the item set as P_{rec}, and the dialogue module predicts the probability distribution over vocabulary as $P_{\text{dial}} \in \mathbb{R}^{|V|}$. The overall probability of generating the next token is calculated as follows:

$$P(w_o) = p_s P_{\text{dial}}(w) + (1 - p_s)\, P_{\text{rec}}(i) \quad (1)$$

$$p_s = \sigma\,(W_s e + b_s) \quad (2)$$

where w_o represents either a word from the vocabulary or an item from the item set, e is the hidden representation in the final layer of the dialogue module. σ refers to the sigmoid function and W_s and b_s are the learnable parameters.

4 Method

In this section, we present the framework of learning **N**eural **T**emplates for **R**ecommender **D**ialogue system, called **NTRD**. As shown in Figure 1, NTRD mainly consists of three components: a recommendation-aware response template generator, a context-aware item selector and a knowledge graph (KG) based recommender. Given the dialogue context, the encoder-decoder based template generator focuses on generating the response template with item slots (Section 4.1). Then the blank slots are filled by the item selector according to the dialogue context, candidate items from the recommender module and the generated response template (Section 4.2). Finally, the entire framework is trained in an end-to-end manner (Section 4.3).

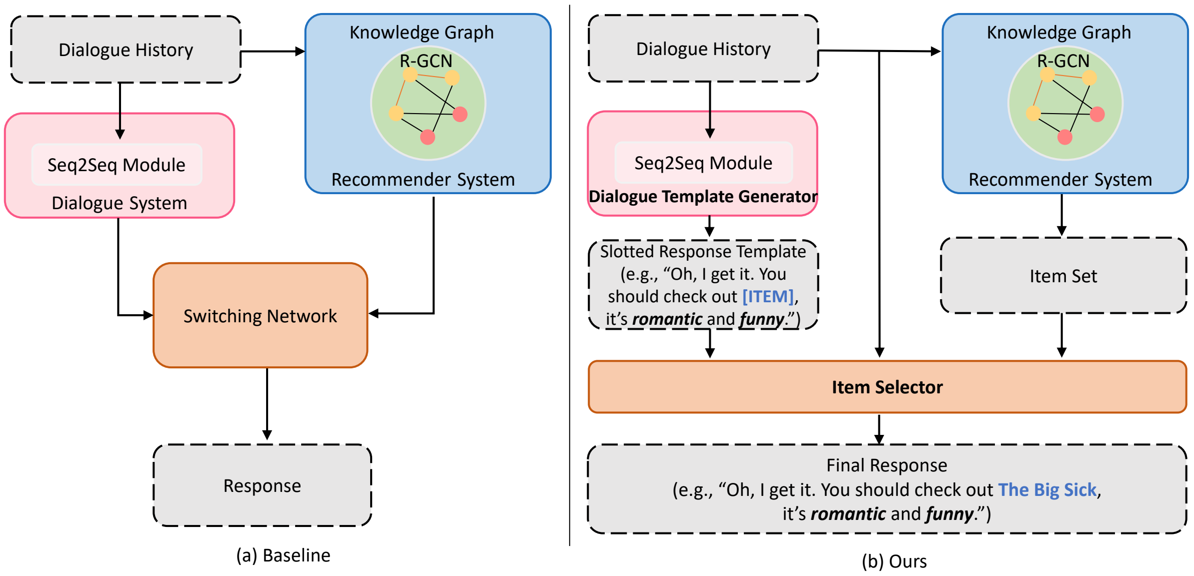
Dialogue History
Knowledge Graph
R-GCN
Recommender System
Seq2Seq Module
Dialogue System
Switching Network
Response
(a) Baseline
Dialogue History
Knowledge Graph
R-GCN
Recommender System
Seq2Seq Module
Dialogue Template Generator
Slotted Response Template (e.g., "Oh, I get it. You should check out [ITEM], it's romantic and funny.")
Item Set
Item Selector
Final Response (e.g., "Oh, I get it. You should check out The Big Sick, it's romantic and funny.")
(b) Ours

Figure 1: Comparison on modules of the existing frameworks and our proposed NTRD framework.

4.1 Response Template Generator

To generate the response template, we adopt the Transformer-based network (Vaswani et al., 2017) to model the process. Concretely, we follow Zhou et al. (2020a) to use the standard Transformer encoder architecture and the KG-enhanced decoder which can effectively inject the information from KG into the generation process. Then we add a special token [ITEM] into the vocabulary and mask all items in the utterances of dialogue corpus with [ITEM] tokens. Thus, at each time step, the response template generator predicts either the special token [ITEM] or the general words from the vocabulary. Formally, the probability of generating the next token by the response template generator is given as follows:

$$P_{\text{dial}}(w) = softmax\left(W_d e + b_d\right) \qquad (3)$$

where $W_d \in \mathbb{R}^{|\mathcal{V}| \times d^e}$ and $b_d \in \mathbb{R}^{|\mathcal{V}|}$ are weight and bias parameters, d^e is the embedding size of the hidden representation e. After the generation process is finished, these special tokens serve as the item slots in generated templates, which will be filled with the specific items by item selector.

4.2 Slot Filling with Item Selector

Now we have the generated response templates, the rest we need to do is filling the slot locations with proper items. Here, we first reuse the KG-enhanced recommender module (Zhou et al., 2020a) to get the user representation given the dialogue context. The recommender module learns a user representation p_u through incorporating two special knowledge graphs, *i.e.*, a word-oriented KG (Speer et al., 2017) to provide the relations between words and an item-oriented KG (Bizer et al., 2009) to provide the structured facts regarding the attributes of items. Given the learned user preference p_u, we can compute the similarity between user and item as follows:

$$similarity\,(m) = softmax\left(p_u^{\mathrm{T}} \cdot h_m\right) \qquad (4)$$

where h_m is the learned embedding for item m, and d^h is the dimension of h_m. Hence, we rank all the items for p_u according to Eq. 4 and produce a candidate set from the total item set.

Existing works (Chen et al., 2019; Zhou et al., 2020a) infer the final item only based on the dialogue context. While the generated response template can also provide the additional information for selecting the final item. For instance, as shown in the example of Figure 1, the words ***"romantic"*** and ***"funny"*** after item slot could provide the contextual semantic information in the response for choosing the item to be recommended.

Motivated by this, we propose a context-aware item selector by stacking sufficient multi-head attention blocks, as shown in Figure 2. Formally, we define the embedding matrix E_{slot} for all the slots in the template, where each slot embedding is the hidden representation from the final layer of transformer decoder. Similarly, the embedding matrix for the remaining tokens in the template is

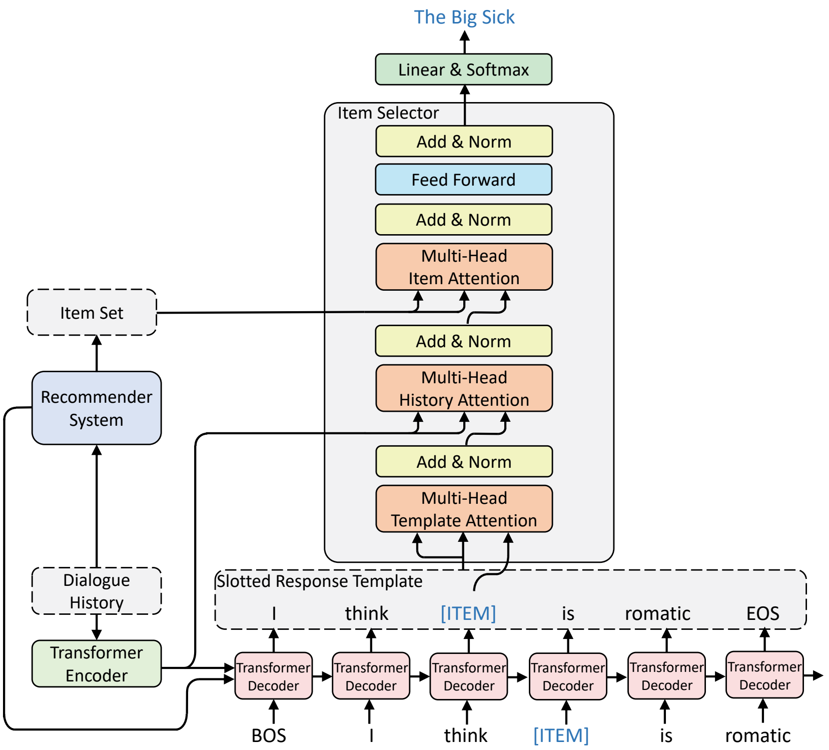

Figure 2: The overview of our approach. The slotted response template is first generated by the transformer decoder and then the item selector fills in the slots with proper items. Our framework enables sufficient information interaction among the generated template, dialogue history, and candidate items in a progressive manner, which is beneficial to selecting the more suitable items to fill in the slot locations.

defined as E_{word} and the embedding matrix output by the transformer encoder is E_{ctx}. H_{cand} is the concatenated embedding matrix for candidate items. Hence, the calculation in the item selector is conducted as follows:

$$\begin{aligned} \hat{E_{slot}} &= \mathrm{MHA}\left(E_{slot}, E_{wrod}, E_{word}\right) \\ E'_{slot} &= \mathrm{MHA}\left(\hat{E_{slot}}, E_{ctx}, E_{ctx}\right) \\ E''_{slot} &= \mathrm{MHA}\left(E'_{slot}, H_{cand}, H_{cand}\right) \end{aligned} \tag{5}$$

where $\mathrm{MHA}(Q, K, V)$ defines the multi-head attention function (Vaswani et al., 2017) that takes a query matrix Q, a key matrix K, and a value matrix V as input and outputs the attentive value matrix:

$$\begin{aligned} &\mathrm{MHA}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_\mathrm{h})W^O, \\ &\text{where head }_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{6}$$

Note that the layer normalization with residual connections and fully connected feed-forward network are omitted in Eq. 5 for simplicity. By this means, the item selector is able to sufficiently fuse effective information from the generated template, dialogue context and candidate items in a progressive manner, which is beneficial to selecting the more suitable items to fill in the slot locations.

Finally, the item selector predicts a probability distribution over all items and selects the one with the highest score to fill in:

$$P_{\text{rec}}(w) = softmax\left(W_r e_{\text{slot}} + b_r\right) \tag{7}$$

where $W_r \in \mathbb{R}^{|\mathcal{M}_t| \times d^e}$ and $b_r \in \mathbb{R}^{|\mathcal{M}_t|}$ are weight and bias parameters.

4.3 Training Objectives

Though the entire framework is typically two-stage, the two modules can be trained simultaneously in an end-to-end manner. For the template generation, we optimize a standard cross-entropy loss as:

$$L_{\text{gen}} = -\sum_{t=1}^{N} \log\left(P_{\text{dial}}(s_t | s_1, ..., s_{t-1})\right) \tag{8}$$

where N is the number of turns in a conversation $\mathcal{D}$, s_t is the t-th utterance of the conversation.

While the loss function for the item selector is calculated as:

$$L_{\text{slot}} = -\sum_{i=1}^{|\mathcal{M}_\mathcal{D}|} \log\left(P_{\text{rec}}(m_i)\right) \tag{9}$$

where $|\mathcal{M}_{\mathcal{D}}|$ is the number of ground truth recommended items in a conversation $\mathcal{D}$.

We combine the template generation loss and the slot selecting loss as:

$$L = \lambda L_{\text{gen}} + L_{\text{slot}} \tag{10}$$

where λ is a weighted hyperparameter.

During the inference, we apply greedy search to decoding the response template $s_t = (w_1, w_2, ..., w_s)$. If w_i is the special token `[ITEM]`, the item selector will be used to select the appropriate specific item based on the dialogue context, generated template and candidate items. Finally, the completed response will be sent to the user to carry on the interaction.

5 Experimental Setup

5.1 Dataset

To evaluate the performance of our method, we conduct comprehensive experiments on the ReDial dataset[1], which is a recent CRS benchmark (Li et al., 2018). This dataset collects high-quality dialogues for recommendations on movies through crowd-sourcing workers on Amazon Mechanical Turk (AMT). It contains 10,006 conversations consisting of 182,150 utterances related to 6,924 movies, which is split into the training, validation, and test set in an 80-10-10 proportion.

5.2 Evaluation Metrics

Both automatic metrics and human evaluation are employed to evaluate the performance of our method. For dialogue generation, automatic metrics include: (1) **Fluency**: perplexity (PPL) measures the confidence of the generated responses. (2) **Diversity**: Distinct-n (Dist-n) (Li et al., 2016) are defined as the number of distinct n-grams divided by the total amount of words. Specifically, we use Dist-2/3/4 at the sentence level to evaluate the diversity of generated responses.

For recommendation task, existing works (Li et al., 2018; Chen et al., 2019; Zhou et al., 2020a) individually evaluate the performance on recommendation using Recall@k. However, the goal of open-ended CRS is to smoothly chat with users and naturally incorporate proper recommendation items into the responses. In other words, it is important for the system to generate informative replies containing the accurate items. Hence, we introduce a new metric that checks whether the ground-truth item is included in the final generated response, *i.e.*, Recall@1 in Response (ReR@1). Similarly, if the generated response has an item token, we calculate whether the top-k (k=10, k=50) items of the probability distribution for this position contain the ground truth item, *i.e.*, ReR@10 and ReR@50. Besides, we also introduce the Item Diversity that measures the percentage of the recommended items mentioned in the generated response to all items in the dataset. Item Ratio is introduced by Zhou et al. (2020a) to measure the ratio of items in the generated response.

For human evaluation, 100 dialogues are randomly sampled from the test set. Then three crowd-workers are employed to score on the generated responses in terms of *Fluency* and *Informativeness*. The range of score is 1 to 3. The higher score means the better. The average score of each metric on these 100 dialogues evaluated by three workers is reported. The inter-annotator agreement is measured by Fleiss' Kappa (Fleiss and Cohen, 1973).

5.3 Implementation Details

The models are implemented in PyTorch and trained on one NVIDIA Tesla V100 32G card. For the fair comparison, we keep the data preprocessing steps and hyperparameter settings the same as the KGSF model (Zhou et al., 2020a) in the released implementation[2]. The embedding size d^h of the item in recommender module is set to 128, and the embedding size d^e in dialogue module is set to 300. We follow the procedure in KGSF to pre-train the knowledge graph in the recommender module using Mutual Information Maximization (MIM) loss for 3 epochs. Then the recommender module is trained until the cross-entropy loss converges. For the training of response template generator, we replace the movies mentioned in the corpus with a special token `[ITEM]` and add it to the vocabulary. We use Adam optimizer with the $1e-3$ learning rate. The batch size is set to 32 and gradient clipping restricts in [0, 0.1]. The generation loss and the item selection loss are trained simultaneously with the weight $\lambda = 5$.

5.4 Baselines

We introduce the baseline models for the experiments in the following:

- **ReDial** (Li et al., 2018): The baseline model

[1] https://redialdata.github.io/website/

[2] https://github.com/RUCAIBox/KGSF

proposed by Li et al. (2018) consists of an auto-encoder (Wang et al., 2015) recommender, a dialogue generation model based on HRED (Serban et al., 2017) and a sentiment prediction model.

- **KBRD** (Chen et al., 2019): This model utilizes a KG to enhance the user representation. The transformer-based (Vaswani et al., 2017) dialogue generation model uses KG information as the vocabulary bias for generation.
- **KGSF** (Zhou et al., 2020a): The model proposes to incorporate two external knowledge graphs, *i.e.*, a word-oriented KG and an item-oriented KG, to further enhance in modeling the user preferences.

6 Experimental Results

6.1 Evaluation on Dialogue Generation

We conduct the automatic and human evaluations to evaluate the quality of generated responses.

Automatic Evaluation. Table 1 shows the automatic evaluation results of the baseline models and our proposed NTRD on dialogue generation. As we can see, our NTRD is obviously better on all automatic metrics compared to the baseline models. Specifically, NTRD achieves the best performance on PPL, which indicates the generator of NTRD can also generate the fluent response templates. In terms of diversity, NTRD consistently outperforms the baselines with a large margin on Dist-2/3/4. This is because the generated template provides the extra contextual information for slot filling so as to produce more diverse and informative responses.

Human Evaluation. We report the human evaluation results in Table 2. All Fleiss's kappa values exceed 0.6, indicating crowd-sourcing annotators have reached the substantial agreement. Compared to KGSF, our NTRD performs better in terms of Fluency and Informativeness. NTRD decouples the response generation and item injection by first learning response templates and then filling the slots with proper items. Hence, it can generate the more fluent and informative responses in practice.

6.2 Evaluation on Recommendation

In this section, we evaluate the performance of recommendation from two aspects, *i.e.*, conversational item recommendation to assess the recall performance and novel item recommendation to investigate the generalization ability.

Conventional Item Recommendation. To further investigate the performance of NTRD on the conventional item recommendation, we present the experimental results of ReR@k (k=1, 10 and 50), Item Diversity and Item Ratio in Table 1. As can be seen, when evaluating the actual performance of recommendation based on final produced responses, the state-of-the-art method KGSF performs poorly with only 0.889% ReR@1. This indicates the switching network in KGSF cannot accurately incorporate the recalled items into the generated responses. It violates the original intention of the open-ended CRS, *i.e.*, to not only smoothly chat with users but also recommend precise items using free-form natural text. By contrast, our NTRD framework performs significantly better, which shows the decoupling strategy brings an obvious advantage of incorporating the precise items into the conversations with users. Furthermore, NTRD achieves the highest item ratio and item diversity. On the one hand, the template generator introduces a special token `[ITEM]` and thus reduces the size of vocabulary, which would increase the predicted probability of item slot during the generation process. On the other hand, the item selector utilizes sufficient information from dialogue context, generated template and candidate items to help select the high-quality recommended items.

Novel Item Recommendation. Existing methods have one major drawback that they cannot handle the novel items never appearing in the training corpus. To validate the unique advantage of our NTRD on novel item recommendation, we conduct an additional experiment. Specifically, we collect all items from the test set that do not appear in the training set, *i.e.*, 373 novel items in total. To learn the representations of these novel items, we first include them together with other ordinary items in the pre-training of the recommender modules of both KGSF and NTRD. However, when training the dialogue modules, we only use the normal training set where these novel items are excluded. Then we evaluate the models on the test set. As we can see in Table 3, the 13.40% (50 of 373) of novel items can be successfully incorporated into the final responses and thus recommended to the users, while KGSF fails to recommend any of the novel items. This verifies the promising generalization ability of NTRD on the unseen items, which is

Model	PPL	Dist-2	Dist-3	Dist-4	ReR@1	ReR@10	ReR@50	Item Diversity	Item Ratio
ReDial (Li et al., 2018)	28.1	0.225	0.236	0.228	-	-	-	-	15.8
KBRD (Chen et al., 2019)	17.9	0.263	0.368	0.423	-	-	-	-	29.6
KGSF (Zhou et al., 2020a)	5.55	0.305	0.466	0.589	0.889	1.083	1.733	6.03	31.5
NTRD (ours)	**4.41**	**0.578**	**0.820**	**1.005**	**1.806**	**12.503**	**31.592**	**11.05**	**66.77**

Table 1: Automatic evaluation results on the ReDial dataset. Numbers in bold denote that the improvement over the best performing baseline is statistically significant.

Method	Fluency	Informativeness	Kappa
KGSF	2.24	1.92	0.67
NTRD (ours)	**2.48**	**2.16**	0.62
Human	2.85	2.37	0.74

Table 2: Human evaluation results on the dialogue generation. Human responses come from the ground truth responses provided in the dataset.

Model	Novel Ratio
KGSF	0.00
NTRD (ours)	**13.40**

Table 3: Evaluation results of the experiments on the ability of predicting the novel item.

crucial in the real-world settings. Even though the recommender modules would update regularly to include novel items, our framework can still effectively incorporate these items into conversational responses without re-training the dialogue model.

6.3 Case Study

In this section, we present an interactive example to demonstrate how our model works in the interactive dialogue. More cases please refer to the Appendix. In Table 4, we mark all the mentioned items in blue color and the user preferences in red color for easy reading. The outputs of three recommenders including human, KGSF and NTRD are presented for the comparison. The dialogue starts with greetings between the user (seeker) and the recommenders, followed by the recommenders proactively seeking user preference by asking which kind of movie he/she likes. With the focused preference of the user on "*comedy*" movies, the recommenders provide some candidate movie items with the interesting comments. The responses of KGSF tend to be boring and dull, and it does not switch its recommendation item even though the user reject it by saying "I saw that one". Instead, our NTRD behaves more like human and shares a similar taste with the human recommender (both recommend the comedy movie "*Shaun of the Dead*"). Then when the user expresses intention to watch, both our NTRD and human recommender respond with goodwill and end the dialogue, while KGSF still tries to give a new movie recommendation. In a nutshell, our NTRD is able to provide informative and interesting item recommendation at the utterance level, and behaves more naturally and interactively at the dialogue level.

Conversation Case	
User:	Hello!
Human: **KGSF**: **NTRD**:	What kind of movies do you like? Hi! What kind of movies do you like? Hello, what type of movies are you looking for?
User:	I love comedy movies like *Blades of Glory*.
Human: **KGSF**: **NTRD**:	I see, you can watch *Shaun of the Dead*. I would recommend *Bad Santa*. How about *Anchorman 2: The Legend Continues*.
User:	I saw that one. A classic parody film.
Human: **KGSF**: **NTRD**:	Also *Hot_Fuzz* is funny. I really though *Ghostbusters* was funny. Or you can watch *Hangover*. I would recommend *Bad Santa*. Have you seen *Shaun of the Dead*? It's a really good one.
User:	I am a big fan of both of those movies as well.
Human: **KGSF**: **NTRD**:	But *Deadpool 2 (2018)* just came out and it 's funny as well. Have you seen it? I also like *Hot Fuzz*. I think you will like *Borat*.
User:	Not yet, but I plan too.
Human: **KGSF**: **NTRD**:	Ok. Hope I helped, bye. I also like *Groundhog Day*. I hope you enjoy your night!

Table 4: A sampled case between a real user as a seeker, and the dialogue agents (including human, KGSF and our NTRD) as recommenders. Items mentioned are marked in the blue color, while the user preferences in user's turn are marked in the red color.

7 Conclusions

In this paper, we introduce **NTRD**, a novel recommender dialogue framework that decouples the dialogue generation from the item recommendation via a two-stage strategy. Our approach reconciles classical slot filling approaches with modern neural NLG approaches, which make the recommender dialogue system more flexible and controllable. Besides, our approach exhibits the promising generalization ability to recommend novel items that do not appear in the training corpus. Extensive experiments show our approach significantly outperforms the previous state-of-the-art methods.

For future work, the generalization ability of NTRD could be further explored. Current method only supports one placeholder with broad semantics to represent all the item mentions in the dialogue corpus, which is lack of fine-grained annotation. One possible attempt is to extend it to support fine-grained item placeholders, such as replacing the placeholder with different attributes of the items, to further improve its performance.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia-a crystallization point for the web of data. *Journal of web semantics*, 7(3):154–165.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, Hong Kong, China. Association for Computational Linguistics.

Yihong Chen, Bei Chen, Xuguang Duan, Jian-Guang Lou, Yue Wang, Wenwu Zhu, and Yong Cao. 2018. Learning-to-ask: Knowledge acquisition via 20 questions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1216–1225. ACM.

Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen. 2020. Towards explainable conversational recommendation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2994–3000. ijcai.org.

Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H. Chi. 2018. Q&r: A two-stage approach toward interactive recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 139–148. ACM.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. *arXiv preprint arXiv:2105.09710*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany. Association for Computational Linguistics.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, Online. Association for Computational Linguistics.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 173–182. ACM.

Huang Hu, Xianchao Wu, Bingfeng Luo, Chongyang Tao, Can Xu, Wei Wu, and Zhan Chen. 2018. Playing 20 question game with policy-based reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3233–3242, Brussels, Belgium. Association for Computational Linguistics.

Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston.

2019. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1951–1961, Hong Kong, China. Association for Computational Linguistics.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020a. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 304–312. ACM.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, Melbourne, Australia. Association for Computational Linguistics.

Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020b. Interactive path reasoning on graph for conversational recommendation. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 2073–2083. ACM.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9748–9758.

Shijun Li, Wenqiang Lei, Qingyun Wu, Xiangnan He, Peng Jiang, and Tat-Seng Chua. 2021. Seamlessly unifying attributes and items: Conversational recommendation for cold-start users. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–29.

Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2019. Deep conversational recommender in travel. *arXiv preprint arXiv:1907.00710*.

Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4):2065–2073.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. Revcore: Review-augmented conversational recommendation. *arXiv preprint arXiv:2106.00957*.

Wenchang Ma, Ryuichi Takanobu, Minghao Tu, and Minlie Huang. 2020. Bridging the gap between conversational reasoning and interactive recommendation. *arXiv preprint arXiv:2010.10333*.

Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. 2020. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6344–6354, Online. Association for Computational Linguistics.

Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Zi Huang, and Kai Zheng. 2021. Learning to ask appropriate questions in conversational recommendation. *arXiv preprint arXiv:2105.04774*.

Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Nguyen Quoc Viet Hung, Zi Huang, and Xiangliang Zhang. 2020. Crsal: Conversational recommender systems with adversarial learning. *ACM Transactions on Information Systems (TOIS)*, 38(4):1–40.

Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000. IEEE.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Rajdeep Sarkar, Koustava Goswami, Mihael Arcan, and John P. McCrae. 2020. Suggest me a movie for tonight: Leveraging knowledge graphs for conversational recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4179–4189, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *arXiv preprint arXiv:1801.01957*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 235–244. ACM.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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

Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 1235–1244. ACM.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xianchao Wu, Huang Hu, Momo Klyen, Kyohei Tomita, and Zhan Chen. 2018. Q20: Rinna riddles your mind by asking 20 questions. *Japan NLP*.

Hu Xu, Seungwhan Moon, Honglei Liu, Bing Liu, Pararth Shah, Bing Liu, and Philip Yu. 2020. User memory reasoning for conversational recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5288–5308, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kerui Xu, Jingxuan Yang, Jun Xu, Sheng Gao, Jun Guo, and Ji-Rong Wen. 2021. Adapting user preference to online feedback in multi-round conversational recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 364–372.

MIAO Yisong. 2020. *Advanced Method Towards Conversational Recommendation*. Ph.D. thesis, National University OF Singapore.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tong Zhang, Yong Liu, Peixiang Zhong, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. Kecrs: Towards knowledge-enriched conversational recommendation system. *arXiv preprint arXiv:2105.08261*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics: System Demonstrations, pages 270–278, Online. Association for Computational Linguistics.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018b. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 177–186. ACM.

Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. 2017. Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 27–36, Saarbrücken, Germany. Association for Computational Linguistics.

Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. CRSLab: An open-source toolkit for building conversational recommender system. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 185–193, Online. Association for Computational Linguistics.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020a. Improving conversational recommender systems via knowledge graph based semantic fusion. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1006–1014. ACM.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020b. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4128–4139, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jie Zou, Yifan Chen, and Evangelos Kanoulas. 2020. Towards question-based recommender systems. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 881–890. ACM.

Jie Zou and Evangelos Kanoulas. 2019. Learning to ask: Question-based sequential bayesian product search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 369–378. ACM.

A Appendix

In this section, we show more conversation examples. We mark all the mentioned items in blue color and the user preferences in red color for easy reading.

Conversation Case	
User:	Hi.
Human:	Tell me what kind of movies would you like to watch?
KGSF:	What kind of movies do you like?
NTRD:	What type of movies do you like?
User:	I like horror and thriller. Especially low budget like *The Boy (2016 film)* and *Don't Breathe*.
Human:	Have you seen *The Babadook*?
KGSF:	I haven't seen either. I heard of that one. I will be *Don't Breathe* and *Insidious: Chapter 2* and *Insidious: Chapter 2*.
NTRD:	Have you seen *Sweeney Todd: The Demon Barber of Fleet Street*?
User:	No, but I want to!
Human:	What about *The Cabin in the Woods*?
KGSF:	I haven't seen either. I heard of that one. I will be *Don't Breathe*.
NTRD:	*Sweeney Todd: The Demon Barber of Fleet Street* is a great one.
User:	Yes, I like that one.
Human:	I really enjoy *The Cabin in the Woods*?
KGSF:	I haven't seen either. I heard of that one. I will be *Don't Breathe*.
NTRD:	Yes, *It (2017 film)* is really good.
User:	Yeah, me too.
Human:	Have you seen *Raw (film)*? It is pretty scary.
KGSF:	I hope you enjoy them!
NTRD:	Yes, *Don't Breathe* is really good.
User:	No, I haven't heard of that. Thanks for the suggestions.
Human:	Bye. You are Welcome.
KGSF:	You 're welcome. Bye.
NTRD:	You are welcome. Enjoy your day.

Table 5: Example (A).

Conversation Case	
User:	User: I like to watch scary movies.
Human:	Tell me what kind of movie do you like ? Did you watch *It (2017 film)*? And *It (1990 film)*?
KGSF:	I would recommend some scary ones, are you like *Scream (1996 film)* or *Annabelle* and *Insidious: Chapter 2*?
NTRD:	What about *The Shining*?
User:	I have seen the old one.
Human:	Ok, you should watch the new one. Even though i would rather the old one. And did you watch *The Last Exorcism*? I hope i have helped.
KGSF:	I would recommend *The Babadook*.
NTRD:	Have you seen *It Follows*? That 's a great one and is a great one of my favorites.
User:	No, I have not but it sounds scary! Thank you for the suggestions!
Human:	You should try them. You 're welcome. Good bye.
KGSF:	Bye.
NTRD:	You 're welcome. Bye.

Table 6: Example (B).

Conversation Case	
User:	Hello , are you a movie buff ?
Human:	Hi, I guess I could say yes from the ridiculous stack of movies :) You?
KGSF:	Hello! What kind of movies do you like?
NTRD:	Hi, what type of movies do you like?
User:	I watch many movies. I especially like science fiction. Any recommendations ?
Human:	An oldy but a goody is *The Wizard of Oz (1939)*. I prefer more actions and comedies, any recomendations?
KGSF:	I haven't seen that one. I 'm glad I could help.
NTRD:	How about *Alien: Covenant (2017)*?
User:	I like Bruce Willis action.
Human:	I really enjoy movies like *The Avengers (2012)* and *Iron Man (2008)* . Recently we watched *Spider-Man: Homecoming (2017)* and it was great!
KGSF:	I haven't seen that one. I will be good. I will be good. I will be
NTRD:	*Casablanca* is good as well.
User:	I like *Seven Samori* as well as *The Magnificent Seven*.
Human:	I like Bruce Willis in a lot of movies, especially in *Die Hard (film series)*.
KGSF:	I haven't seen that one. I will be good. I will be good. I will be
NTRD:	*Iron Man (2008)* was good too.
User:	I will watch *The Avengers(2012)* again , Thanks!
Human:	I will add it to the Netflix que. Thank you!
KGSF:	Have a great day!
NTRD:	You too.

Table 7: Example (C).