Target-guided Knowledge-aware Recommendation Dialogue System: An Empirical Investigation

Dongding Lin[1], Jian Wang[1] and Wenjie Li[1]

[1]Department of Computing, The Hong Kong Polytechnic University

Abstract

The target-guided recommendation dialogue system aims to make high-quality recommendations through interactive conversations proactively and naturally. Existing methods still struggle to incorporate background knowledge for coherent response generation, and to recommend appropriate items with respect to dialogue context, user preference and recommendation target. In this paper, we investigate the problem of target-guided knowledge-aware recommendation dialogue and design a dialogue generation system to alleviate the above-mentioned issues. Specifically, we employ pre-trained language models with multi-task learning to jointly learn response generation and goal prediction towards the target. We also present a knowledge-preserving encoding strategy to maintain the facts in background knowledge. Extensive experiments on two benchmark datasets show that our system significantly outperforms various competitive models in terms of both automatic and manual evaluations. We further provide analysis and discussions to demonstrate that our system is effective in leveraging both related knowledge and planned goals to generate fluent, informative and coherent responses towards the target of recommendation.

Keywords

Recommendation Dialogue, Background Knowledge, Target Guiding, Multi-task Learning

1. Introduction

Building a human-like dialogue system is one of the long-cherished goals in natural language processing (NLP) [1]. Dialogue systems can be mainly used for chatting with users for entertainment, i.e., open-domain dialogues [2, 3], or accomplishing specific tasks, i.e., task-oriented dialogues [4, 5, 6]. Recent years, recommendation dialogue systems [7, 8] have been recognized as an important special type of task-oriented dialogue systems with the aim of discovering user preferences and making recommendations through conversations. The growing research interests mainly come from the benefits that dialogue provides an effective channel to handle the cold-start problem in recommendations while recommendation-oriented tasks promotes technological advance for dialogue systems [9].

Many existing methods have focused on various aspects of both recommendation and conversation, including user preference modeling [10], conversation strategy [11, 12], and dialogue generation [13, 14]. Today, incorporating knowledge graphs (KG) has been recognized as crucial to bridge the gap between recommendation reasoning and response generation [13]. Background knowledge helps a dialogue system better understand user interests and make recommendations via coherent natural language communication. Since users may not have a clear preference for the unfamiliar new items recommended, especially in many sociable recommendation domains such as music, movies and news, it is also important for a dialogue system to proactively lead the conversation to the recommendation target with high user engagement and enjoyment.

Recently, the emergence of the DuRecDial [15] dataset provided new insights towards the development of target-guided knowledge-aware recommendation dialogue systems. As the example shown in Figure 1, the whole user-bot dialogue is grounded on a user profile, background knowledge, and a goal sequence. The bot needs to take both the user's interests and the knowledge graph into consideration to decide an optimal goal path to achieve the target of recommending and playing a music (i.e, "*Days of Friendship*"). Here, the goal path is a sequence of goals, with each goal specifying a goal type (e.g., "*Greeting*" or "*Movie Recommendation*") and a goal topic (e.g., the movie "*Orphans of the Zhao Family*" or the music "*Ode to Joy*"). It is challenging to build such an effective recommendation dialogue system, as the planning of goals which is required to ensure a natural and smooth conversation must take into consideration of both the user interests and the ultimate target goal. The joint modeling of dialogue context, user profile and recommendation target remains under-explored in the existing work.

In this paper, we investigate two key challenges in recommendation dialogue systems: (1) how to effectively

3rd Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) & 5th Edition of Recommendation in Complex Environments (ComplexRec) Joint Workshop @ RecSys 2021, September 27–1 October 2021, Amsterdam, Netherlands

csdlin@comp.polyu.edu.hk (D. Lin);
csjiwang@comp.polyu.edu.hk (J. Wang);
cswjli@comp.polyu.edu.hk (W. Li)

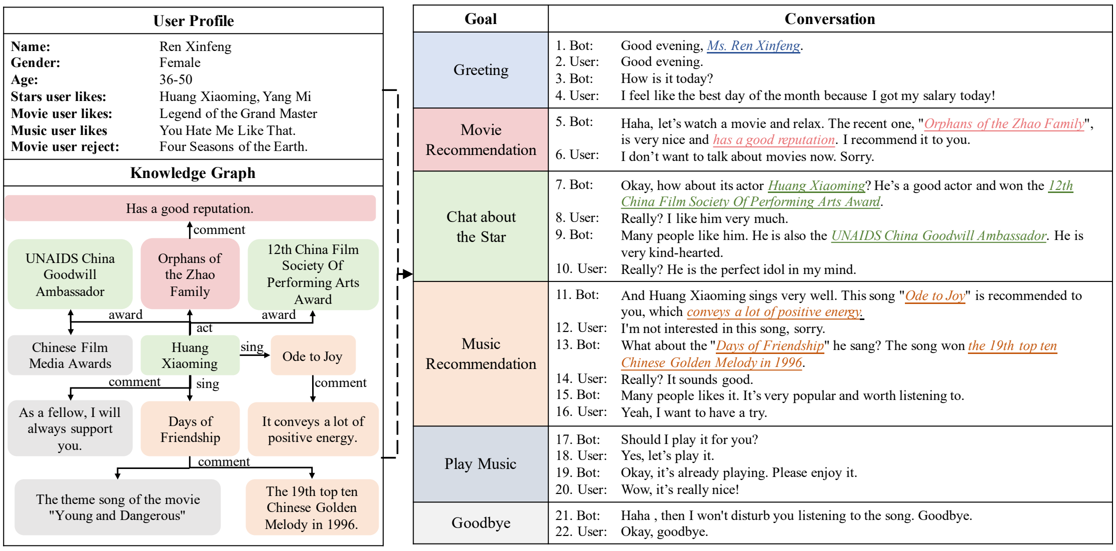

Figure 1: An illustrative example of target-guided knowledge-aware recommendation dialogue from the DuRecDial dataset. The whole dialogue is grounded on user profile, background knowledge, and a goal sequence, where the goal sequence is planned by the bot to achieve the target of recommending and playing a music.

incorporate related facts in background knowledge and user profile in dialogue generation, and (2) how to make appropriate goal planning to proactively lead the conversation. To this end, we employ multi-task learning to jointly learn response generation and goal prediction towards the final target based on pre-trained language models. Specifically, we adopt ERNIE-GEN [16], an enhanced multi-flow pre-training and fine-tuning framework for natural language generation, as our backbone model. In addition, we also present a knowledge-preserving encoding strategy to maintain the background knowledge facts for dialogue generation. Extensive experiments on two benchmark datasets show that our system significantly outperforms various competitive models in terms of both automatic and manual evaluations. We have submitted our best model to Baidu Language and Intelligence Challenge 2021 (LIC 2021[1]), where we achieved the 4-th rank among 862 teams. It reveals that our methods are effective to generate informative, coherent and appropriate responses and to achieve the target of recommendation.

Overall, our contributions are three folds:

(1) Towards building a target-guided recommendation dialogue system, we adopt multi-task learning to jointly model goal planning and dialogue generation based on pre-trained language models.

(2) We present a knowledge-preserving encoding strategy to better maintain background knowledge facts in order to enhance the system ability to generate appropriate responses by incorporating background knowledge.

(3) The evaluation results show that our system achieves significant improvement compared to various competitive models.

[1]https://aistudio.baidu.com/aistudio/competition/detail/67?isFromLuge=true

2. Related Work

The two research lines that motivate our study are conversational recommendation systems (CRS) and recommendation dialogue systems. We briefly introduce some representative works as below.

2.1. Conversational Recommendation System

A conversational recommendation system (CRS) is a recommendation system (RS) that provides personalized recommendation through natural language conversations. Christakopoulou et al. [17] argued that asking questions benefit a RS , which can better understand user preference based on user feedback. To this end, they suggested to move from the traditional RS to the CRS. Lei et al. [11] proposed a three-stage framework called Estimation-Action-Reflection (EAR) to fill the interaction gap between conversation and recommendation in implicit ways. More explicitly, Lei et al. [12] leveraged conversational recommendation as finding a path in a user-item-attribute graph interactively. To enhance semantic representations of products and related textual descriptions of products, both Zhou et al. [18] and Rajdeep et al. [19] incorporated external knowledge graphs (KG) into CRS, which in turn led to better recommendations. However, despite the improvement towards high-quality recommendations, these methods have limited abilities to generate natural and informative dialogues.

2.2. Recommendation Dialogue System

A recommendation dialogue system is a special type of task-oriented dialogue system, which is expected to encourage natural human-machine interaction with a clear target. To facilitate the research along this line, several recommendation dialogue datasets have been released, including GoRecDial [8] and INSPIRED [20]. To further investigate whether the system can lead a multi-type dialogue to approach the target of recommendation with rich interaction behavior, Liu et al. [15] created a large-scale dialogue dataset, namely DuRecDial. Existing recommendation dialogue approaches mainly focus on how to effectively integrate interactive recommendation and dialogue generation. Cai et al. [21] contributed two hierarchical taxonomies for classifying user intents and recommendation actions. To bridge the gap between recommendation reasoning and response generation, Ma et al. [13] performed tree-structured reasoning on knowledge graphs, which can then be mapped to hierarchical dialogue acts to guide generation. More recently, Bai et al. [14] proposed a goal-oriented knowledge copy network to discern the knowledge facts that are highly correlated to the dialogue, which assisted to generate accurate knowledge-aware responses.

In this paper, we aim to build a target-guided dialogue system towards recommendation. It requires the system to make high-quality recommendations by considering external knowledge and user preference. More importantly, the system should also be able to lead the conversation towards the target goal naturally by generating appropriate responses.

3. Method

3.1. Problem Definition

Suppose a target-guided dialogue corpus is denoted as $D = \{(H_i, K_i, G_i, P_i, Y_i)\}_{i=1}^{N}$, where $H_i = \{h_{i,t}\}_{t=1}^{T}$ represents dialogue history with multiple turns, $K_i = \{k_{i,j}\}_{j=1}^{N_K}$ is a set of background knowledge facts that correspond to this conversation and each element $k_{i,j}$ is formulated as a triplet. $G_i = \{g_{i,j}\}_{j=1}^{L_G}$ is a goal sequence which is constructed upon the knowledge facts set K_i and each goal $g_{i,j}$ consists of a goal type and a goal topic. $P_i = \{p_{i,j}\}_{j=1}^{N_P}$ represents a set of user profiles with each profile $p_{i,j}$ in the format of $\langle key, value \rangle$ pair. $Y_i = \{y_{i,j}\}_{j=1}^{L_Y}$ is the response produced on the basis of the H_i, K_i, G_i, and P_i. Here, L_G and L_Y denote the sequence length of G_i and Y_i respectively.

Given explicit goals $G' = \{g_1, g_{L_G}\}$ (i.e., start goal and target goal), a dialogue history H' paired with the related knowledge facts K', and the user profile P', the objective of target-guided recommendation dialogue is to decide an appropriate goal g_c at each turn to determine where the dialogue should go with the aim of proactively leading the dialogue from the start goal to the target goal, and meanwhile generating a coherent and informative response to achieve the goal g_c.

3.2. Model Architecture

Backbone Model To tackle the issue of proactively planning goals for target-guided recommendation dialogue, we jointly model goal prediction and dialogue generation based on pre-trained language models, as shown in Figure 2 (a). Pre-trained language models have been widely used in dialogue generation on a basis of pre-training fine-tuning framework, where they generally concatenate different information sources such as knowledge facts and dialogue history as input, and generate responses autoregressively. In this paper, we employ ERNIE-GEN [16], an enhanced multi-flow pre-training and fine-tuning framework for natural language generation, as our backbone framework. ERNIE-GEN bridges the discrepancy between training and inference with an infilling generation mechanism using multi-flow attention. In light of the fact that entities and phrases are organized span by span, ERNIE-GEN adopts span attention mask matrices (see Figure 2 (b)) to determine whether each token and each span can attend to each other. To better capture coherent semantic information of the context, both word-by-word flow and span-by-span flow are integrated together (see Figure 2 (c)), where the the span-by-span generation flow aims to predict semantically complete spans consecutively. In view of the fact that specific entities or spans (e.g., musics, movies, and news) should be generated in the response as the recommended items, we believe ERNIE-GEN is a good choice with the advantages described above.

Knowledge-preserving Encoding It is difficult for existing pre-trained models to encode concatenated background knowledge facts because it often exceeds the encoding length limitation of the models. In particular, according to our statistics, the concatenated background knowledge facts of each dialogue in the DuRecDial [15] dataset contains more than 1,700 tokens on average. It substantially exceeds the encoding length limitation (i.e.,

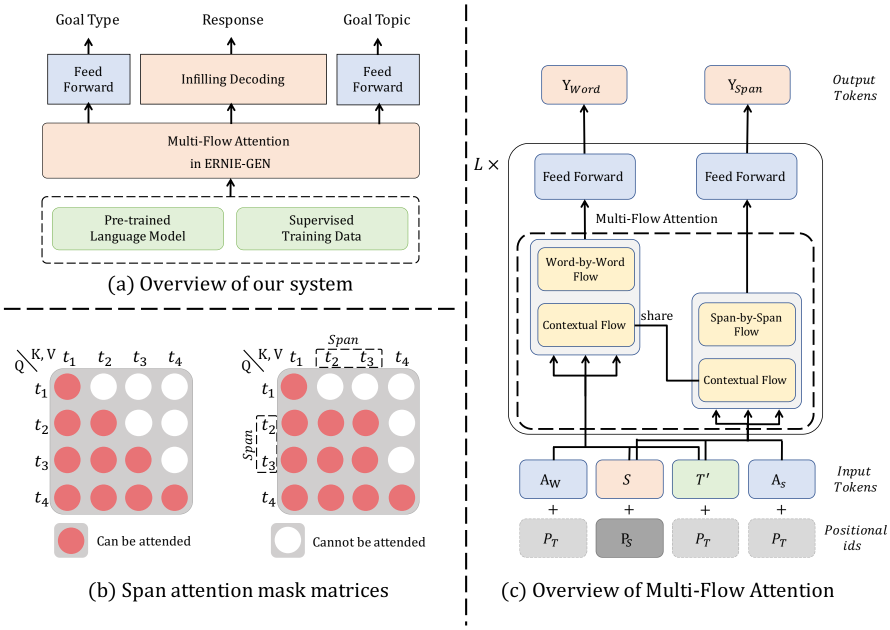

Figure 2: Illustration of methodology. (a): Overview of our system, with goal type prediction, goal topic prediction, and response generation jointly modeled in a multi-task learning manner. (b): The span attention mask matrices used in our system. (c): Overview of Multi-Flow attention in ERNIE-GEN [16], which is employed as our backbone model.

512) of many pre-trained language models including ERNIE-GEN [16]. To address this issue, we present a knowledge-preserving encoding strategy to better maintain background knowledge facts. First, all knowledge triplets K_i in i-th dialogue are sorted according to the token length of each triplet after being concatenated. Then, we put these knowledge triplets into a number of buckets $\{B_{i,j}\}_{j=1}^{N_{B_i}}$ with the short-length-first-in priority, where $B_{i,j} \subseteq K_i$, N_{B_i} denotes the number of buckets. The capacity of each bucket $B_{i,j}$ is tuned by a hyper-parameter C, which denotes that $B_{i,j}$ contains no more than C tokens in total. We hope that after concatenating $B_{i,j}$ with other information sources, the total input length fulfills the encoding length limitation (i.e., 512). To this end, the i-th dialogue sample is split into N_{B_i} dialogue samples. Note that the system will generate multiple responses during inference with this strategy. We adopt a simple unsupervised strategy to select the "best" one. We calculate mutual F1 scores by treating one response as the "ground-truth" and the others as the candidate generated results. The average F1 score of the candidate results will be regarded as the selection score for the "ground-truth". Therefore, each response will obtain a corresponding selection score. We select the response with the highest selection score as the final generated response.

Multi-task Learning As described in Section 3.1, the system should generate a coherent and informative response following an appropriate goal, which is decided by the system itself at each turn with the aim of proactively leading the dialogue from the start goal to the target goal. Intuitively, the goal planning process has an important effect on dialogue generation. To this end, we propose to add goal prediction at each turn as an auxiliary task, which are jointly fine-tuned with the dialogue generation task in a multi-task learning manner. Concretely, we divide the task of goal prediction into two sub-tasks, goal type prediction and goal topic prediction. We feed the hidden representation of ERNIE-GEN's encoding output to two individual fully-connected feed forward neural networks, followed by a softmax operator, both of which are optimized using cross-entropy loss. As shown in Figure 2 (a), the fine-tuning objective during the training stage is to jointly optimize the goal type prediction loss $\mathcal{L}_{type}$, goal topic prediction loss $\mathcal{L}_{topic}$, and the response generation loss $\mathcal{L}_{gen}$. We minimize the following overall loss:

$$\mathcal{L} = \beta_1 \mathcal{L}_{type} + \beta_2 \mathcal{L}_{topic} + \mathcal{L}_{gen} \tag{1}$$

where β_1, β_2 are two hyper-parameters controlling the impact of the goal type and the goal topic. Under the supervision of goal planning in the training stage, the system will learn to naturally generate coherent responses so as to achieve goal transition during the inference stage.

4. Experiments

4.1. Datasets

We conduct extensive experiments on two knowledge-aware recommendation dialogue datasets, i.e., DuConv [22] and DuRecDial [15] that are accompanied with the explicitly specified goals. We also use some other dialogue datasets to enhance the fine-tuning process. All datasets are in Chinese.

- **DuConv:** It consists of about 30k dialogues and 270k utterances in the movie domain. Each dialogue contains about 14 background knowledge triplets on average. The goal sequence of each dialogue is an explicit path "$[start] \rightarrow topic\_a \rightarrow topic\_b$" over the knowledge graph, indicating how a dialogue is led from any start point relevant to $topic\_a$ to the final $topic\_b$. Here, $topic$ represents one entity in the background knowledge.
- **DuRecDial:** It is composed of about 10k dialogues and 156k utterances over multi-type domains, including chit-chat, question answering (QA), and music/movie/news recommendation, etc. Each dialogue session consists of about 15 turns on average, with about 22 background knowledge triplets and a specified user profile (e.g., age, gender, preference) in the format of $\langle key, value \rangle$ pairs. The goal sequence is constructed upon the knowledge and user profiles, with each goal containing a goal type and a goal topic (entity). There are altogether 21 goal types.
- **Other Datasets:** Since it is important to select appropriate entities or phrases from background knowledge facts for a recommendation dialogue system, we also utilize additional large-scale dialogue datasets to help fine-tune our system because of their similar settings for incorporating knowledge in dialogue generation. The datasets include ESTC [23], Tencent [24], and KdConv [25]. Both ESTC and Tencent datasets are collected from open-domain conversations, with about 900k and 5.5M dialogues respectively. The KdConv dataset covers conversations about movie, music, and tourism, which has more than 3k dialogues. We will discuss the effect of model performance with and without using these datasets in Section 5.1.

4.2. Data Preprocessing

To better understand the characteristics of different datasets, we conduct data analysis and preprocessing first. The statistics of DuConv and DuRecDial datasets are reported in Table 1. For the DuConv dataset, it has an average of 4.5 dialogue turns and an average of 14.2 knowledge triplets. After concatenating the multi-turn dialogue history, goal topics, knowledge triplets as well as the input sequence, we observe that it fulfills the maximum input length (i.e., 512) of ERNIE-GEN in most cases. For those few samples that exceed the length limitation, we simply take the last 512 tokens (i.e., Chinese characters) as input. For the DuRecDial dataset, it has an average of about 15 dialogue turns and an average of about 22 knowledge triplets. After taking user profiles and goal sequences into account, the average length of each concatenated input sequence substantially exceeds the maximum input length (i.e., 512) of ERNIE-GEN. Therefore, we adopt the knowledge-preserving encoding strategy described in Section 3.2 to better maintain the background knowledge facts for each dialogue.

4.3. Baselines

We compare our system with baseline models and several competitive methods as follows.

- **Seq2Seq**[2] [26] is a generative baseline used in many dialogue generation tasks. We concatenate dialogue history, knowledge facts, and other sources (if any) together as the input sequence and feed it to the vanilla sequence-to-sequence (Seq2Seq) model with the attention mechanism to generate responses.
- **MGCG_R/G** [15] include a retrieval-based model and a generation-based model for multi-goal driven conversation generation. They are presented as the second baseline on the DuRecDial dataset.
- **UniLM** [27] is a unified pre-trained language model that can be used for language generation by controlling generation with specific self-attention masks.
- **GPT-2** [28] is an autoregressive pre-trained language model and has been successfully used in many downstream language generation tasks. The pre-training on large-scale text corpora makes it easy to be fine-tuned for dialogue generation.
- **GOKC**[3] [14] is a generation-based model with a goal-oriented knowledge discernment mechanism, which discerns the knowledge facts that are highly correlated to the dialogue goal and the dialogue context. Note that GOKC is the publicly available state-of-the-art model on both the DuConv dataset and the DuRecDial dataset.

4.4. Implementation Details

Our dialogue system is built on top of the official open-source code of ERNIE-GEN[4]. During training (fine-tuning), both β_1 and β_2 are set to 1.0 and the batch size

[2]https://opennmt.net/OpenNMT-py/.
[3]https://github.com/jq2276/Learning2Copy
[4]https://github.com/PaddlePaddle/ERNIE/tree/repro/ernie-gen

Table 1
The statistics of DuConv dataset and DuRecDial dataset

			Goals	History				Response		KB triplets	
		#Dialogue	Avg. size	Max. turn	Avg. turn	Max. length	Avg. length	Max. length	Avg. length	Max. size	Avg. size
DuConv	Train	17,858	2	9	4.5	354	17.1	130	21.3	23	14.2
	Dev	2,000	2	9	4.5	77	17.0	77	21.2	21	14.2
	Test	2,000	2	9	4.5	100	17.1	77	22.3	21	14.2
DuRecDial	Train	6,018	4.5	28	15.2	255	16.1	85	22.3	71	21.4
	Dev	600	4.5	26	15.2	250	16.1	79	22.3	56	22.4
	Test	946	4.6	26	15.3	245	16.1	68	22.3	57	21.7

is set to 8. We use Adam [29] optimizer with the initial learning rate of 1×10^{-4}, the L_2 weight decay of 0.01 and the learning rate warm-up over the first 10% training steps with linear decay. During generation, we adopt beam search decoding algorithm with a beam size of 5. The details are described below.

Fine-tuning We start fine-tuning from the pre-trained Chinese version of ERNIE 1.0 model [30], as it is compatible with the ERNIE-GEN framework and its pre-trained model checkpoint can be directly loaded. We first fine-tune our system on 3 large-scale dialogue datasets (as described in Section 4.1) for 5 epochs. Due to the large size of the ESTC dataset and the Tencent dataset, we randomly extract 400K dialogue samples from each original dataset. We continue to fine-tune our system on the target dialogue datasets (DuConv and DuRecDial) for 10 epochs, with the bucket capacity C setting to 360.

Vocabulary Expansion We find that ERNIE-GEN may generate unknown words (i.e., `[UNK]`), i.e., the words out of the vocabulary. Therefore, we add the additional tokens with high occurrence extracted from the datasets to expand the original vocabulary. The final vocabulary size is 18,000, which can cover almost all the Chinese characters and common special tokens in the datasets.

Deduplication We observe that our system tends to generate repeated words or phrases sometimes, which is a common issue that is still under exploration in natural language generation. To make the generated response look more fluent, we remove the consecutive repeated words using regular expression rules.

4.5. Evaluation Metrics

Automatic Evaluation Following the common practice [15, 14], we adopt the following automatic evaluation metrics.

- **F1 score:** It indicates whether the model can generate appropriate entities in the response.
- **BLEU-1/2 scores:** They are also calculated at the character level, representing 1-gram and 2-gram overlaps between the generated response and the gold response.
- **Distinct (DIST)-1/2 scores:** They are used to evaluate the 1-gram diversity and 2-gram diversity of the generated response.
- **Perplexity (PPL):** It is widely used to estimate how well a probability model predicts a sample. A low perplexity indicates the model is good at predicting the sample.

Human Evaluation With human evaluation, we randomly select 100 dialogue samples from the testset, and then invite 5 evaluators to independently assign the rating score for the output of each model following the metrics suggested in [15]. The score of each metric is ranged from 0 to 2. Furthermore, we also report the human evaluation in the Baidu LIC 2021, where crowed-sourcing annotators are invited to conduct about 10 multi-round conversations with each submitted system and to judge the dialogue quality. The metrics used in our evaluation and in Baidu LIC 2021 are in consistent, including:

- **Informativeness (Info.):** It measures if the model makes full use of knowledge facts in the generated response.
- **Coherence (Cohe.):** It measures the overall fluency of the whole dialogue generation.
- **Knowledge accuracy (Know Acc.):** It evaluates the accuracy of the selected knowledge in the generated response.
- **Recommendation success rate (Rec. Succ.):** It estimates how well the target recommendation goal is achieved.

5. Results and Analysis

5.1. Automatic Evaluation

The automatic evaluation results on the DuConv dataset and the DuRecDial dataset are reported in Table 2.

Table 2

The automatic evaluation results on the DuConv dataset and the DuRecDial dataset. The *norm* and *ext* stand for *normalized* and *external dialogue datasets* respectively.

Model	F1	BLEU-1/2	DIST-1/2	PPL
DuConv				
norm retrieval	34.73	0.291/0.156	**0.118/0.373**	N.A.
Seq2Seq	39.94	0.283/0.186	0.093/0.222	10.96
norm generation	41.84	0.347/0.198	0.057/0.155	24.30
UniLM	41.10	0.326/0.213	0.089/0.241	10.05
GPT-2	41.05	0.349/0.223	0.095/0.250	8.72
GOKC	45.09	0.410/0.272	0.105/0.272	9.92
Ours (w/o *ext*)	45.15	0.443/0.361	0.097/0.276	7.81
Ours (w/ *ext*)	**45.72**	**0.450/0.369**	0.106/0.285	**6.49**
DuRecDial				
Seq2Seq	26.08	0.188/0.102	0.006/0.013	22.82
UniLM	29.05	0.226/0.149	0.030/0.096	18.65
MGCG_R	33.93	0.340/0.232	0.068/**0.187**	N.A.
MGCG_G	36.81	0.323/0.219	0.017/0.052	17.69
GPT-2	47.01	0.392/0.295	0.055/0.165	15.56
GOKC	47.28	0.413/0.318	0.025/0.084	11.38
Ours (w/o *ext*)	48.62	0.475/0.401	0.060/0.168	4.42
Ours (w/ *ext*)	48.73	0.476/0.403	0.065/0.171	4.21
Ours (multi-task)	48.80	0.479/0.408	0.064/0.166	4.22
Ours (full-goal)	**48.86**	**0.479/0.410**	**0.070**/0.185	**4.20**

Table 3

The human evaluation results on the DuConv dataset and the DuRecDial dataset.

Model	Info.	Cohe.	Know Acc.	Rec. Succ.
DuConv				
norm retrieval	0.605	0.767	0.677	0.71
norm generation	0.881	0.819	0.904	0.85
UniLM	0.909	1.022	0.921	0.93
GPT-2	0.911	1.216	0.934	0.94
GOKC	1.052	1.308	1.108	1.12
Ours (w/o ext)	1.060	1.436	1.127	1.26
Ours (w/ ext)	**1.081**	**1.456**	**1.213**	**1.31**
DuRecDial				
MGCG_R	1.118	1.260	0.955	0.75
MGCG_G	1.218	1.222	0.985	0.88
UniLM	0.989	1.078	0.852	1.0
GPT-2	1.289	1.256	1.165	1.20
GOKC	1.311	1.258	1.189	1.23
Ours (w/o ext)	1.358	1.262	1.286	1.21
Ours (w/ ext)	1.430	1.262	1.295	1.23
Ours (multi-task)	**1.433**	**1.268**	**1.315**	**1.32**

Our model outperforms all the compared models, and achieves a significant improvement over most of the evaluation metrics. Specifically, on the DuConv dataset, the normalized models (i.e., norm retrieval and norm generation) refer to using normalized data by replacing the specific two goals in the knowledge path with "topic_a" and "topic_b" respectively, following [22]. As shown in Table 2, our model yields substantial improvement over existing pre-trained models including UniLM and GPT-2 on both F1 and BLEU-1/2. It demonstrates that our model can generate more coherent and informative responses in the n-gram's level. Compared to the state-of-the-art model GOKC, our model without using external dialogue datasets (w/o ext) still achieves about 0.13%, 8%, and 32% improvements in terms of F1, BLEU-1, and BLEU-2, respectively. After using external dialogue datasets (w/ ext), our model further achieves 1.39%, 9.7%, and 35.6% improvements of F1, BLEU-1, and BLEU-2 compared to GOKC, which indicates that fine-tuning on large-scale task relevant dialogue datasets is effective to improve the performance in the final target-guided knowledge-aware recommendation dialogue task. Note that the normalized retrieval method achieves the highest DIST-1/2 scores. The retrieval-based methods that directly select responses from a list of candidates is more likely to retain the diversity of the natural responses.

As shown in Table 2, our model also achieves superior performance than all baseline methods on the DuRecDial dataset. In particular, compared to the competitive model GOKC, our model without using external dialogue datasets obtains about 2.8%, 15%, and 26.1% improvements on F1, BLEU-1, and BLEU-2 evaluation metrics. When using external dialogue datasets, our model achieves about 3%, 15.3%, and 26.7% improvements accordingly. Our model with multi-task learning further outperforms baseline methods on all metrics, which demonstrates that joint modeling of goal planning and dialogue generation is effective to help the system select the appropriate knowledge from the background facts to facilitate generation. Besides, we also observe that the perplexity of our model is much lower, indicating that our model is more likely to generate fluent responses. It should be noted that based on its default setting, GOKC actually assumes that the full goal sequence is provided and thus does not require any goal planning [15]. Therefore, for fair comparison we also report our evaluation results using the available full goal sequence in Table 2. The results further show the effectiveness of multi-task learning for our system. Overall, our model achieves significant improvements over competitive methods in terms of all automatic evaluation metrics.

5.2. Human Evaluation

The human evaluation results of baseline models and our model are presented in Table 3. As shown in Table 3, our model obtains the highest human scores on both the DuConv and DuRecDial datasets, which shows the effectiveness to generate informative and coherent responses with correct knowledge and consistent information. Specifically, our model achieves significant improvement in terms of knowledge accuracy, which further verifies that fine-tuning on large-scale task relevant dialogue datasets is effective to improve the ability of our model to incorporate knowledge into generation. We observe that our model with multi-task learning obtains

Table 4
The human evaluation results on the Baidu LIC 2021.

Rank	Info.	Cohe.	Know Acc.	Rec. Succ.
DuConv				
Team-1	1.023	1.469	1.215	N.A.
Team-2	1.109	1.572	1.312	N.A.
Team-3	1.198	1.649	1.412	N.A.
Team-4 (Ours)	1.081	1.456	1.213	N.A.
Team-5	0.881	1.266	0.944	N.A.
Team-6	1.088	1.153	1.176	N.A.
DuRecDial				
Team-1	0.481	1.322	0.795	2.0
Team-2	0.426	1.09	0.641	1.633
Team-3	0.444	1.131	0.664	1.5
Team-4 (Ours)	0.431	1.26	0.706	1.0
Team-5	0.48	1.246	0.754	1.467
Team-6	0.349	1.027	0.533	1.167

Table 5
The experimental results of different implementation details.

Our Model	F1	BLEU-1/2	DIST-1/2	PPL
DuConv				
Base	43.47	0.412/0.294	0.097/0.265	7.12
+ Deduplication	43.60	0.415/0.298	0.099/0.269	7.12
+ knowledge-preserving enc.	44.23	0.422/0.336	0.104/0.276	6.98
+ Vocabulary expansion	44.50	0.435/0.347	0.105/0.279	6.65
+ All	**45.72**	**0.450/0.369**	**0.106/0.285**	**6.49**
DuRecDial				
Base	45.97	0.447/0.389	0.053/0.159	4.55
+ Deduplication	46.02	0.449/0.387	0.057/0.161	4.55
+ knowledge-preserving enc.	47.65	0.458/0.397	0.065/0.164	4.40
+ Vocabulary expansion	48.27	0.463/0.405	0.067/0.179	4.34
+ All	**48.86**	**0.479/0.410**	**0.070/0.185**	**4.20**

much better recommendation success rate on the DuRecDial dataset. This verifies that our joint modeling of goal prediction and response generation enables the system to make more accurate recommendations with respect to the given goals and the user profile.

We submitted our best model to the Baidu LIC 2021 and achieved the 4-th rank among 862 teams. The human evaluation results on the leaderboard are shown in Table 4. Note that the human evaluation here is more challenging due to two aspects: (1) The decision of the current goal relies on the previously predicted goals, and (2) the generation of the response at the current turn will be further decided by the current goal. It is likely to cause error accumulations for a model during multi-turn conversations. Therefore, the evaluation results can better reveal the abilities of different models to guide the conversation to the target. As shown in Table 4, our system is competitive compared to others. However, our model performs inferior on the DuRecDial dataset in terms of the recommendation success rate. It encourages us to further improve goal planning strategies in future work.

5.3. Discussion

Analysis of Implementation Details We study the contribution of each part in our system by conducting experiments with several variants of our system. The results are shown in Table 5. Here is our findings. (1) All strategies are effective to improve dialogue generation performance. (2) The knowledge-preserving encoding strategy contributes significantly to dialogue generation especially on the DuRecDial dataset where the input source sequence is much longer. Compared to previous methods that truncate tokens when the sequence exceeds the encoding length limitation of pre-trained models, our proposed encoding strategy better maintains the knowledge facts. (3) After expanding the vocabulary, our model achieves significant improvements over most of the metrics, which means that the vocabulary containing more common tokens is important in dialogue generation.

Future Research Direction In real world, recommending new target items that possibly attract users is meaningful since users often have no definite preference for many unknown items. We are trying to achieve this objective through the development of the target-guided knowledge-aware recommendation dialogue system. We understand that it is not sufficient by simply modeling the target and dialogue with multi-task learning as investigated in this paper. We will leave the problem of proactively planning goals step by step towards the target goal as our future research direction.

6. Conclusion

In this paper, we explore target-guided knowledge-aware recommendation dialogue based on the pre-training fine-tuning framework, which aims to proactively lead the conversation and learn to make high-quality recommendations. We present a knowledge-preserving encoding strategy and a multi-task learning approach to enable our system to effectively recommend appropriate items and to generate fluent and coherent responses. The experimental results on two benchmark datasets demonstrate the effectiveness and superiority of our system compared to the other competitive models in terms of both automatic and manual evaluations. We also discuss the implementation details and our future research direction.

Acknowledgments

The work described in this paper was supported by Research Grants Council of Hong Kong (PolyU/15207920, PolyU/15207821), National Natural Science Foundation of China (61672445, 62076212) and PolyU Internal Grants (ZVVX, ZG7H, ZVQ0).

References

[1] A. M. Turing, Computing machinery and intelligence, in: Parsing the turing test, 2009, pp. 23–65.

[2] W. Chen, J. Chen, P. Qin, X. Yan, W. Y. Wang, Semantically conditioned dialog response generation via hierarchical disentangled self-attention, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3696–3709.

[3] Y. Wu, F. Wei, S. Huang, Y. Wang, Z. Li, M. Zhou, Response generation by context-aware prototype editing, in: The Thirty-Third AAAI Conference on Artificial Intelligence, 2019, pp. 7281–7288.

[4] M. Eric, L. Krishnan, F. Charette, C. D. Manning, Key-value retrieval networks for task-oriented dialogue, in: Proceedings of the 18th Annual SIG dial Meeting on Discourse and Dialogue, 2017, pp. 37–49.

[5] A. Madotto, C.-S. Wu, P. Fung, Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers), 2018, pp. 1468–1478.

[6] C. Wu, R. Socher, C. Xiong, Global-to-local memory pointer networks for task-oriented dialogue, in: 7th International Conference on Learning Representations (ICLR), 2019.

[7] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, J. Tang, 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), 2019, pp. 1803–1813.

[8] D. Kang, A. Balakrishnan, P. Shah, P. Crook, Y.-L. Boureau, J. Weston, 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), 2019, pp. 1951–1961.

[9] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Computing Surveys (CSUR) 54 (2021) 1–36.

[10] H. Xu, S. Moon, H. Liu, B. Liu, P. Shah, P. S. Yu, User memory reasoning for conversational recommendation, in: Proceedings of the 28th International Conference on Computational Linguistics (COLING), 2020, pp. 5288–5308.

[11] W. Lei, X. He, Y. Miao, Q. Wu, R. Hong, M. Kan, T. Chua, Estimation-action-reflection: Towards deep interaction between conversational and recommender systems, in: J. Caverlee, X. B. Hu, M. Lalmas, W. Wang (Eds.), The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM), 2020, pp. 304–312.

[12] W. Lei, G. Zhang, X. He, Y. Miao, X. Wang, L. Chen, T. Chua, Interactive path reasoning on graph for conversational recommendation, in: R. Gupta, Y. Liu, J. Tang, B. A. Prakash (Eds.), The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2020, pp. 2073–2083.

[13] W. Ma, R. Takanobu, M. Tu, M. Huang, Bridging the gap between conversational reasoning and interactive recommendation, arXiv preprint arXiv:2010.10333 (2020).

[14] J. Bai, Z. Yang, X. Liang, W. Wang, Z. Li, Learning to copy coherent knowledge for response generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 12535–12543.

[15] Z. Liu, H. Wang, Z.-Y. Niu, H. Wu, W. Che, T. Liu, Towards conversational recommendation over multi-type dialogs, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 1036–1049.

[16] D. Xiao, H. Zhang, Y. Li, Y. Sun, H. Tian, H. Wu, H. Wang, ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), 2020, pp. 3997–4003.

[17] K. Christakopoulou, F. Radlinski, K. Hofmann, Towards conversational recommender systems, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 815–824.

[18] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J. Wen, J. Yu, Improving conversational recommender systems via knowledge graph based semantic fusion, in: R. Gupta, Y. Liu, J. Tang, B. A. Prakash (Eds.), The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2020, pp. 1006–1014.

[19] R. Sarkar, K. Goswami, M. Arcan, J. P. McCrae, Suggest me a movie for tonight: Leveraging knowledge graphs for conversational recommendation, in: Proceedings of the 28th International Conference on Computational Linguistics (COLING), 2020, pp. 4179–4189.

[20] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, Z. Yu, INSPIRED: Toward sociable recommendation dialog systems, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8142–8152.

[21] W. Cai, L. Chen, Predicting user intents and satisfaction with dialogue-based conversational rec-

ommendations, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 33–42.
[22] W. Wu, Z. Guo, X. Zhou, H. Wu, X. Zhang, R. Lian, H. Wang, Proactive human-machine conversation with explicit conversation goals, arXiv preprint arXiv:1906.05572 (2019).
[23] H. Zhou, M. Huang, T. Zhang, X. Zhu, B. Liu, Emotional chatting machine: Emotional conversation generation with internal and external memory, in: S. A. McIlraith, K. Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 730–739.
[24] D. Cai, Y. Wang, W. Bi, Z. Tu, X. Liu, S. Shi, Retrieval-guided dialogue response generation via a matching-to-generation framework, 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), 2019, pp. 1866–1875.
[25] H. Zhou, C. Zheng, K. Huang, M. Huang, X. Zhu, KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 7098–7108.
[26] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, 2014, pp. 3104–3112.
[27] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, H. Hon, Unified language model pre-training for natural language understanding and generation, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS), 2019, pp. 13042–13054.
[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[29] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations (ICLR), 2015.
[30] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, H. Wu, Ernie: Enhanced representation through knowledge integration, arXiv preprint arXiv:1904.09223 (2019).