

KERS: A Knowledge-Enhanced Framework for Recommendation Dialog Systems with Multiple Subgoals

Jun Zhang¹, Yan Yang^{1,2*}, Chengcai Chen³, Liang He^{1,2}, Zhou Yu⁴

¹East China Normal University

²Shanghai Key Laboratory of Multidimensional Information Processing

³Xiaoi Research, Xiaoi Robot Technology Co., Ltd

⁴Columbia University

51194506048@stu.ecnu.edu.cn, {yanyang, lhe}@cs.ecnu.edu.cn,

arlenecc@xiaoi.com, zhouyu@cs.columbia.edu

Abstract

Recommendation dialogs require the system to build a social bond with users to gain trust and develop affinity in order to increase the chance of a successful recommendation. It is beneficial to divide up, such conversations with multiple subgoals (such as social chat, question answering, recommendation, etc.), so that the system can retrieve appropriate knowledge with better accuracy under different subgoals. In this paper, we propose a unified framework for common knowledge-based multi-subgoal dialog: **Knowledge-enhanced multi-subgoal driven recommender system (KERS)**. We first predict a sequence of subgoals and use them to guide the dialog model to select knowledge from a sub-set of existing knowledge graph. We then propose three new mechanisms to filter noisy knowledge and to enhance the inclusion of cleaned knowledge in the dialog response generation process. Experiments show that our method obtains state-of-the-art results on *DuRecDial* dataset in both automatic and human evaluation.

1 Introduction

Recommendation dialog systems recently attract much attention due to their significant commercial potential (Chen et al., 2019; Jannach et al., 2020). Such systems first elicit user preferences through conversations and then provide high-quality recommendations based on elicited preferences.

Many real-world recommendation applications usually involve chitchat, question answering, and recommendation dialogs working together (Wang et al., 2014; Ram et al., 2018). Various social interactions build rapport with users and gain trust. To provide more sociable recommendations, Liu et al. (2020) proposed a conversational recommendation dialog dataset *DuRecDial* annotated with 21 subgoals, where the dialog system starts the

conversation with some non-recommendation subgoals, such as chitchat and question answering to collect user information and build social relationships and finally progresses into a recommendation subgoal. Subgoals can be seen as different dialog phases. Figure 1 shows an example dialog with multiple subgoals. All the subgoals are designed to complete the final recommendation.

An RNN-based **multi-goal driven conversation generation framework (MGCG)** was proposed to address this task by Liu et al. (2020). MGCG first models the subgoals separately to plan appropriate subgoal sequences for topic transitions and final recommendations. Then MGCG extracts knowledge features from the whole knowledge graph and produces responses to complete each subgoal. However, MGCG did not investigate how to effectively use knowledge in different subgoals. As shown in Figure 1, a conversation often involves a relatively large knowledge graph and multiple subgoals. Both the question answering and the recommendation processes require assistance from accurate knowledge information. Therefore, having rich and accurate knowledge is essential in generating engaging conversations. Since taking all possible knowledge as input will lead to more noise and high computation, how to select useful knowledge in different subgoals is important.

We propose KERS to use knowledge effectively in multi-subgoal conversational recommendation tasks. In order to control the flow of the conversation, we develop a dialog guidance module that predicts a sequence of subgoals and selects useful external knowledge information with respect to each subgoal to improve generation performance. In addition, we propose a sequential attention mechanism, a noise filter, and a knowledge enhancement module to make generated responses more informative. Specifically, the sequential attention mechanism enhances subgoal guidance, the noise filter eliminates unrelated and unnecessary knowledge,

* Corresponding author

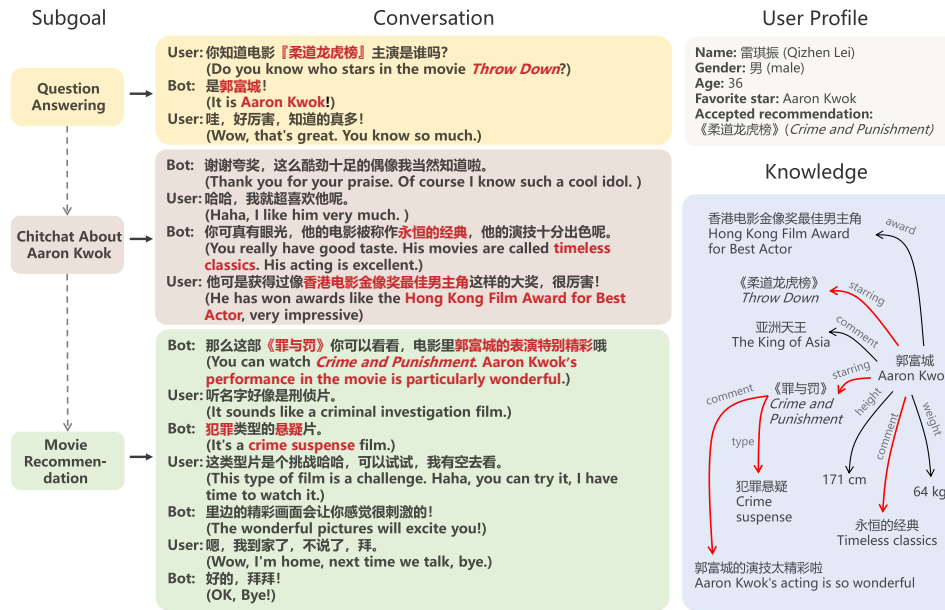


Figure 1: An example of rich knowledge in multi-subgoal recommendation dialog. The conversation is grounded on a knowledge graph. The task can be viewed as completing multiple subgoals sequentially. Text in red indicates knowledge related information and red arrows indicate selected knowledge triple.

and the knowledge enhancement module increases the importance of the selected knowledge in response generation. Both automatic and manual evaluations suggest that KERS has a better performance compared to state-of-the-art methods.

2 Related Work

Most previous work in recommendation dialog systems focused on slot-filling methods to collect user preferences and recommend items (Reschke et al., 2013; Christakopoulou et al., 2016; Sun and Zhang, 2018; Christakopoulou et al., 2018; Zhang et al., 2018; Lee et al., 2018; Lei et al., 2020). To study more sociable and informative recommendation conversations, Li et al. (2018); Moon et al. (2019); Zhou et al. (2020b) proposed new recommendation dialog datasets with knowledge graphs, and incorporated knowledge into response generation. Kang et al. (2019) created a dialog dataset with clear goals. Chen et al. (2019) captured knowledge-grounded information and used recommendation-aware vocabulary bias to improve the quality of language generation.

Recently, Liu et al. (2020) proposed utilizing subgoal sequences to plan dialog paths and presented a new recommendation dialog dataset *DuRecDial*. They demonstrated that establishing a subgoal sequence is crucial for natural transitions and successful recommendations. Some previous works (Moon et al., 2019; Tang et al., 2019; Wu

et al., 2019; Zhou et al., 2020b) also introduced topic transition approaches similar to the subgoal transition to improve the quality of open-domain dialogs. They built the topic path by either traversing on a knowledge graph or predicting knowledge items directly. Similar to Liu et al. (2020), Hayati et al. (2020) utilized sentence-level sociable recommendation strategy labels in the INSPIRED dataset to improve the recommendation success rate. However, the INSPIRED dataset was not annotated with specific dialog subgoals.

Some relevant works for our project focused on obtaining knowledge information from all the related knowledge triples (Liu et al., 2020; Chen et al., 2019), or enhancing the semantic representations by incorporating both word-oriented and entity-oriented knowledge graphs (Zhou et al., 2020a). However, our work differs because it has fine-grained knowledge planning and accurate knowledge incorporation in generation. Moreover, we deal with more complex knowledge graphs, including both sentences and entities.

3 Method

KERS consists of three modules: a dialog guidance module (section 3.1), an encoder (section 3.2), and a decoder (section 3.3), as shown in Figure 2. The decoder incorporates three new mechanisms, a sequential attention mechanism, a noise filter, and a knowledge enhancement module.

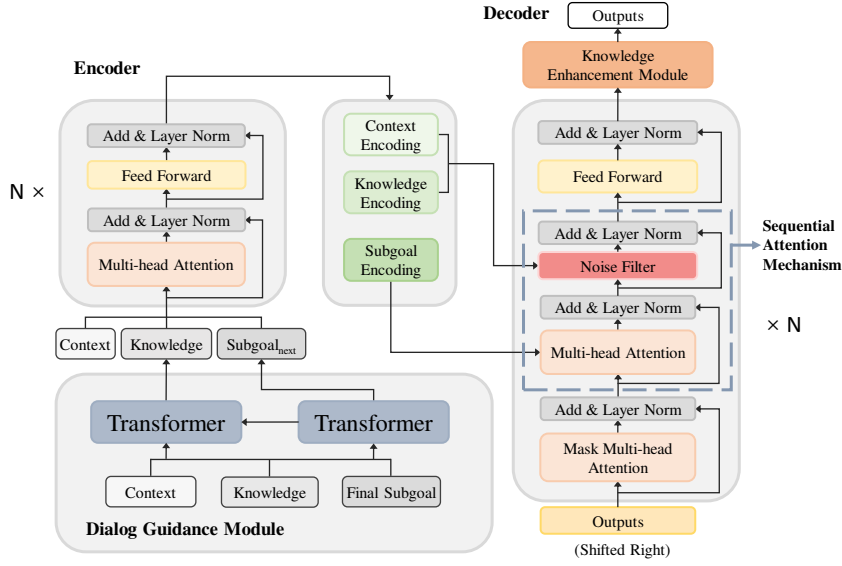


Figure 2: The architecture of the knowledge-enhanced multi-subgoal driven recommender system (KERS).

For each conversation turn, the dialog guidance module predicts the subgoal of the turn and selects dialog knowledge for the next response. Then, the encoder encodes the subgoal, the selected knowledge, and the dialog context. Finally, the output of the encoder is fed to the decoder to generate the final dialog system response.

3.1 Dialog Guidance Module

To produce proactive and natural conversational recommendations, we propose a dialog guidance module to customize a reasonable sequence of subgoals and provide proper candidate knowledge. This module accomplishes two subtasks: subgoal generation and knowledge generation. To predict the next turn’s subgoal G_{next} , we use a Transformer (Vaswani et al., 2017) based model conditioning on a context X , a knowledge graph \mathcal{K} , a user profile \mathcal{P} , and a final recommendation subgoal G_T . We define \mathcal{K}' as a set of \mathcal{P} and \mathcal{K} , and optimize the following loss function:

$$L_G = \sum_i -\log P(g_i^{next} | X, \mathcal{K}', G_T, g_{<i}^{next}) \quad (1)$$

where g_i^{next} denotes the token in G_{next} . Then we input the predicted subgoal into another Transformer to get the candidate knowledge K_c . Because there is no labeled knowledge in ground-truth responses, we obtain pseudo labels in an unsupervised manner. We first concatenate the knowledge items in the tuple $(head, relation, tail)$. Then we compute the char-based F1 score (Wu et al., 2019)

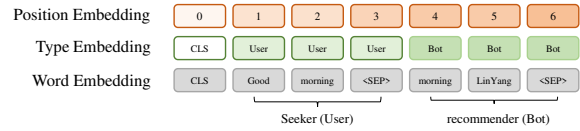


Figure 3: Input representation of the dialog context.

between each knowledge and the ground-truth response. Finally, we take the knowledge items with F1 scores greater than a threshold ($thr = 0.35$) as the pseudo label K_w . We optimize the following loss function to train a knowledge generator:

$$L_K = \sum_i -\log P(k_i^w | G_{next}, X, \mathcal{K}', G_T, k_{<i}^w)$$

where k_i^w is the token in *head* or *relation*. We do not need to generate a complete tuple $(head, relation, tail)$, because only *head* and *relation* are needed to obtain specific knowledge items. Then, we select the knowledge items matching the generated tuple $(head, relation)$ as the candidate knowledge K_c . Finally, the dialog guidance module outputs $G'_{next} = [G_{next}; G_T]$ (the concatenation of the predicted subgoal G_{next} and the final recommendation subgoal G_T) and K_c for next stage processing.

3.2 Encoder

To incorporate different types of information, we use a vanilla Transformer block as our encoder. We encode context, candidate knowledge selected and

the subgoals predicted by the dialog guidance module independently, since they have different structures. In addition, the input embedding includes word embedding, type embedding, and positional embedding, as shown in Figure 3. The multi-type embeddings help the encoder distinguish different parts of the context better (Wolf et al., 2018). Formally, the outputs of the encoder are computed as follows:

$$E_C = \text{Transformer}(X) \quad (2)$$

$$E_K = \text{Transformer}(K_c) \quad (3)$$

$$E_G = \text{Transformer}(G'_{next}) \quad (4)$$

3.3 Decoder

We propose three new mechanisms to incorporate in a Transformer based decoder to generate informative responses consistent with the predicted subgoal. We describe the three mechanisms, a sequential attention mechanism, a noise filter, and a knowledge enhancement module in details below. The decoder produces responses as follows:

$$Y = \arg \max_{Y'} P(Y'|E_C, E_K, E_G) \quad (5)$$

3.3.1 Sequential Attention Mechanism

The sequential attention mechanism is designed to enhance subgoal guidance by simulating human cognitive process. Humans first form an overall idea of a recommendation and then pitch the recommendation given the current conversation context. So we make the decoder first processes the different parts of the encoder outputs at different layers and then combine these layers in a particular order that resembles human cognition. Specifically, the Transformer based decoder extracts features as follows:

$$O_P = \text{MultiHead}(I(Y_p), I(Y_p), I(Y_p)) \quad (6)$$

$$O_G = \text{MultiHead}(O_P, E_G, E_G) \quad (7)$$

$$O_{KG} = \text{NF}(O_G, E_C, E_K) \quad (8)$$

$$O_{dec} = \text{FFN}(O_{KG}) \quad (9)$$

where $\text{MultiHead}(Q, K, V)$ is the multi-head attention operation described in Vaswani et al. (2017). Y_p is the previous decoded tokens. $I(\cdot)$ is the embedding function of the input and $\text{NF}(\cdot)$ indicates the process of the noise filter. In this structure, the model captures valid information in the context and the knowledge based on the subgoals and then generates more coherent responses that are consistent with these subgoals.

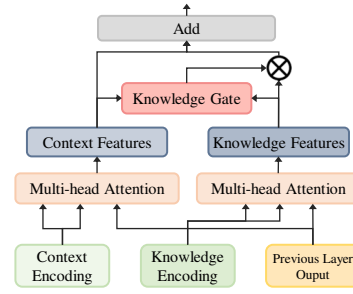


Figure 4: The internal structure of the noise filter.

3.3.2 Noise Filter

Although we can generate high-quality candidate knowledge, there is still erroneous candidate knowledge that can lead to an unexpected response. Moreover, since the recommender does not always provide knowledge-related responses in conversations, the excessive input of knowledge can create more noise. To address these problems, we propose a noise filter to select better knowledge items, shown in Figure 4. We filter the knowledge features by a knowledge gate. Specifically, the filter first takes the previous layer output O_G as a query to extract the features of context encoding E_C and knowledge encoding E_K by multi-head attention:

$$O_C = \text{MultiHead}(O_G, E_C, E_C) \quad (10)$$

$$O_K = \text{MultiHead}(O_G, E_K, E_K) \quad (11)$$

Then, the knowledge gate computes a reduction weight α_k according to the matching degree of knowledge and context. Finally, the filter averages context features and knowledge features using $\alpha_k \in [0, 1]$ as outputs O_{KG} :

$$\alpha_k = \text{Sigmoid}(W_k[O_C; O_K]) \quad (12)$$

$$O_{KG} = O_C + (1 - \alpha_k)O_C + \alpha_k O_K \quad (13)$$

where W_k is a trainable parameter. The noise filter controls the flow of knowledge. When responses are not knowledge-related, or the knowledge is not associated with the context, the reduction weight α_k decreases and vice versa.

3.3.3 Knowledge Enhancement Module

To further generate more informative responses, we propose a knowledge enhancement module to put more emphasis on retrieved knowledge through a set of learned weights. Specifically, we take the words in knowledge \mathcal{K}' as the knowledge lexicon. Then we compute the weighted probability distri-

Model	Accuracy
CNN (Liu et al., 2020)	94.13
LSTM-CNN	95.48
Ours	96.60

Table 1: Subgoal prediction accuracy.

butions of words using a weight $\alpha_g \in [0, 1]$:

$$\alpha_g = \text{Sigmoid}(W_g O_{dec}) \quad (14)$$

$$H = W_v O_{dec} \quad (15)$$

$$P_o(y_j) = \text{Softmax}\left(\begin{bmatrix} \alpha_g H(y_j \notin \mathcal{K}') \\ H(y_j \in \mathcal{K}') \end{bmatrix}\right) \quad (16)$$

where W_g and W_v are trainable parameters. α_g controls the weight of generating a general word. A low value of α_g indicates highlighting the words in the knowledge lexicon. In the training process, the model automatically learns to enhance the generation probability of the knowledge words at proper steps. The introduced knowledge enhancement module can not only help the model produce more informative responses but also increase the presence of the selected knowledge in responses.

3.4 Training Objective

Because that each module completes different functions, we train the model in two stages. First, we optimize the subgoal generation loss L_G and the knowledge generation loss L_K for the dialog guidance module. Then, we optimize the following cross-entropy loss between the predicted word distribution P_o and ground-truth distribution o :

$$L_{RG} = -\sum_{j=1}^N o_j \log(P_o(y_j)) \quad (17)$$

4 Experiments

4.1 Dataset and Training Details

DuRecDial is a dataset for recommendation dialog with annotated subgoals (Liu et al., 2020) in Mandarin. Two crowd workers are assigned different profiles in the recommendation task with a diverse set of subgoals. There are four main categories of subgoals: 1) Chitchat: greeting, chitchat about celebrities, etc; 2) Question answering: answering questions on weather, celebrities, movies, restaurants, music, time, etc; 3) Recommendation: recommending movies, news, music, restaurants, etc; 4) Task: requesting news, playing music,

delivering weather reports. *DuRecDial* contains 10,190 recommendation dialogs, 21 subgoals and 222,198 knowledge triples. We split the dataset into train/dev/test data with a ratio of 6.5:1:2.5. Figure 1 shows an example dialog.

We implement KERS in PyTorch¹. Both the encoder and decoder contain six Transformer blocks. Each Transformer block uses 12 attention heads. The word embedding and hidden state sizes are both set to 768. We use a similar encoder-decoder structure that is used for generating responses to accomplish the subgoal generation and knowledge generation task. The vocabulary size is 30,000. The maximum context length is 768.

4.2 Baseline Models

We compare KERS against several baselines:

- **S2S+kg**: We implement the seq2seq model as described in Vinyals and Le (2015) with the attention mechanism and concatenate all the related knowledge and the context as its input.
- **Trans.**: We implement the Transformer model as introduced by Vaswani et al. (2017).
- **Trans.+kg**: We use a knowledge encoder to extract knowledge features. We concatenate knowledge features and the context as the Transformer model’s input.
- **MGCG_G, MGCG_R**: We use the generation and retrieval models based on the MGCG framework introduced by Liu et al. (2020).

To validate the effectiveness of each component, we conduct ablation studies as follows: (1) **KERS w/o DiaGuidance**: without the dialog guidance module; (2) **KERS w/o Subgoal**: without subgoal information input in the decoder; (3) **KERS w/o CandidateKnow**: without the candidate knowledge input in the decoder; (4) **KERS + Topic**: without the candidate knowledge but with the predicted topic as described in Liu et al. (2020); (5) **KERS w/o NoiseFilter**: without the noise filter; (6) **KERS w/o KnowEnhance**: without the knowledge enhancement module; (7) **KERS + Reverse**: KERS first extracts context and knowledge features, then extracts subgoal features; (8) **KERS + Monolayer**: using the monolayer attention mechanism; (9) **KERS + AllKnowledge**: with all the related knowledge rather than the candidate knowledge.

¹Code will be available at <https://github.com/z562/KERS>.

Model	PPL	F1	BLEU-1	BLEU-2	DIST-2	Know- ledg F1	Train Time (minute)
S2S + kg	24.75	24.52	0.1649	0.0792	0.0131	8.37	27
Trans.	9.78	41.79	0.3925	0.2883	0.0502	27.76	44
Trans. + kg	9.40	44.73	0.4192	0.3180	0.0554	31.82	46
MGCG_R	-	33.93	-	0.2320	0.1870	-	-
MGCG_G ²	16.51	36.02	0.3403	0.2351	0.0574	23.67	30
KERS	8.34	50.47	0.4629	0.3619	0.0790	39.03	50
KERS w/o DiaGuidance	8.80	47.51	0.4371	0.3378	0.0812	35.10	37
KERS w/o Subgoal	8.76	48.95	0.4496	0.3514	0.0821	37.98	46
KERS w/o CandidateKnow	8.58	49.61	0.4550	0.3554	0.0751	37.01	43
KERS + Topic	8.40	49.40	0.4529	0.3532	0.0761	37.07	45
KERS w/o NoiseFilter	8.44	48.98	0.4523	0.3522	0.0765	38.27	54
KERS w/o KnowEnhance	8.56	49.21	0.4544	0.3549	0.0682	37.82	49
KERS + Reverse	8.45	49.42	0.4564	0.3562	0.0787	37.90	50
KERS + Monolayer	8.41	49.40	0.4562	0.3563	0.0789	37.98	47
KERS + AllKnowledge	8.50	49.20	0.4507	0.3515	0.0782	36.73	105

Table 2: Response generation results with automatic evaluation metrics on *DuRecDial* test set.

Moreover, we perform automatic evaluations on two subtasks: subgoal generation and knowledge generation. We compare KERS against: (1) **CNN**: the CNN (Kim, 2014) model used in Liu et al. (2020); (2) **LSTM-CNN**: adding LSTM (Hochreiter and Schmidhuber, 1997) before CNN.

4.3 Automatic Evaluation Metrics

We evaluate the models on the original *DuRecDial* test set. We use perplexity (PPL), F1 (Liu et al., 2020), BLEU (Papineni et al., 2002), and DISTINCT (DIST-2) (Li et al., 2016) for common automatic evaluation. Perplexity and DISTINCT measure the fluency and the diversity of generated responses, respectively. F1 and BLEU measure the similarity between the generated responses and ground truth. In addition, we compare the training time (minutes/epoch) for efficiency. We propose a knowledge F1 score to evaluate selected knowledge’s accuracy. Knowledge F1 is the F1 score computed between the generated response and the pseudo label (aka K_w described in Section 3.1). To evaluate two subtasks, we compute subgoal prediction accuracy and knowledge prediction accuracy.

5 Experimental Results

We first evaluate the effectiveness of subgoal prediction and knowledge prediction. Table 1

²Since MGCG_R is a retrieval-based model and has poor results, we mainly compare our model with MGCG_G.

shows subgoal prediction accuracy. Our model achieves the best performance on subgoal prediction (96.60%) compared to CNN and LSTM-CNN. In addition, our model achieves relatively high accuracy 75.6% on knowledge prediction, which serves a solid base to guide response generation.

We present the response generation results in Table 2. Our model, KERS achieves a significant improvement over previous work MGCG_G in perplexity (PPL) by -8.17, F1 +14.45, BLEU-1 +0.1226, BLEU-2 +0.1268, DIST-2 +0.0216, and knowledge F1 +15.36. Notably, KERS has the lowest perplexity and highest knowledge F1, indicating it has the best fluency and knowledge. Due to the advantages of the retrieval model, MGCG_R has high DIST-2, which suggests MGCG_R has more diverse responses. We also conduct an ablation study to evaluate each component’s contribution to KERS’s performance. Results show that after removing the dialog guidance module, KERS’s performance decreases sharply. This suggests that the dialog guidance module plays a crucial role by providing reasonable subgoals and selecting proper knowledge later. Moreover, removing the predicted subgoals leads to worse performance but higher DIST-2. However, after careful inspection of responses generated by KERS w/o Subgoal, we find that these diverse responses are largely irrelevant to the current scene. Therefore, even though these responses are more diverse, they do not lead to suc-

Model	Turn-level results				Dialog-level results			
	Fluency	Appro.	Infor.	Proactivity	Rec. Success	Coher.	Engag.	
Trans. + kg	2.950	2.585	2.075	2.355	3.982	3.700	3.417	
MGCG_G	2.900	2.360	1.945	2.390	3.850	3.017	2.750	
KERS	2.955	2.840	2.420	2.445	4.439	4.150	3.700	

Table 3: Human evaluation results at different levels. The turn-level evaluation uses a 3-point Likert scale and dialog-level evaluation uses a 5-point Likert scale. * refers to a p-value < 0.05 and ** refers to a p-value < 0.01.

Pref. (%)	Trans. + kg	MGCG_G	KERS
Trans. + kg	-	68.3	38.3
MGCG_G	31.7	-	21.7
KERS	61.7	78.3	-

Table 4: Pair-wise preference of the three models

successful recommendations. We also find that using turn-level candidate knowledge boosts knowledge F1 compared to using subgoal-level topics. This is because turn-level candidate knowledge provides more fine-grained information, which guides response generation. Although our knowledge prediction has a relatively high accuracy of 75.6%, there are still 24.4% incorrect cases – some of them do not need knowledge, and some of them receive the wrong knowledge. The noise filter is designed to address these cases, which improves all the metrics, especially improving F1 by 3.0%. In addition, we find removing the knowledge enhancement module sharply decreases KERS’s DIST-2. We also observe the sequential attention mechanism performs better than both the reverse attention and monolayer structure. This indicates that a reasonable attention sequence enables the model to utilize subgoals and knowledge information better. Furthermore, KERS has better results than KERS+AllKnowledge, especially improving knowledge F1 by 6.3%, and only requires half of its training time. This suggests that rather than improving performance, incorporating all the knowledge introduces noise and leads to more training time. Our model can filter unnecessary information and is more efficient and effective.

6 Human Evaluation

Automatic metrics evaluate the model on several specific aspects, while humans can give a holistic evaluation. We conduct human evaluations on both turn level and dialog level to compare three models,

KERS, MGCG_G, and Trans.+kg. In addition, we run a pair-wise preference test among these models.

6.1 Turn-level Evaluation

We randomly sample 200 examples from the test set and let each model generate a response according to a given context, related knowledge graph, and the final recommendation subgoal. We present the generated responses to five human evaluators. They assess the responses in terms of fluency, appropriateness, informativeness, and proactivity using a 3-point Likert scale.

The results are shown in the left portion of Table 3. The inter-rater annotation agreement is measured using the Fleiss’s kappa (Fleiss and Cohen, 1973). The Fleiss’s kappa for fluency, appropriateness, informativeness, and proactivity is 0.81, 0.76, 0.77, and 0.60, respectively. Our model outperforms all the baselines, especially on appropriateness and informativeness. This indicates that KERS can generate more appropriate and informative responses. Moreover, we find both MGCG_G and KERS obtain relatively higher scores than Trans.+kg on proactivity, suggesting that providing subgoal planning is vital in guiding dialogs.

6.2 Dialog-level Evaluation

We ask human evaluators to have conversations directly with the models through an interactive interface. Since there are 21 different subgoals with different requirements and a large number of different subgoal sequences, we have to train evaluators with the guidelines to effectively evaluate the models. Because such training is time-consuming and requires high proficiency, we recruited ten professional evaluators to perform the evaluation instead of recruiting crowd workers. To make sure evaluators can cover a wide range of different conversation contexts, we ask each evaluator to interact with the models in 6 different scenarios sampled

KERS	MGCG_G	Trans. + kg
<p>User:今天几号了?(What's the date today?)</p> <p>Bot: 今天是2018年11月13日。(Today is Nov. 13, 2018.)</p> <p>User:你真棒。谢谢你啊。(OK, thank you.)</p> <p>Bot: 今天也是黄晓明的生日哦。(Today is also Xiaoming Huang's birthday.)</p> <p>User:今天啊,我偶像啊。(Today! He is my idol.)</p> <p>Bot:你眼光真好,他是第10届华语电影传媒大奖最具人气男演员呢!(You have good taste, He is the most popular actor at the 10th Chinese Film Media Awards.)</p> <p>User:他还是联合国艾滋病规划署中国亲善大使呢!(He's also the UNAIDS Goodwill Ambassador for China!)</p> <p>Bot:是啊,那你要不要看看他的电影《撒娇女人最好命》.电影里的佳品,很不错哦。(Yes, do you want to watch his movie <i>Women Who Know How to Flirt Are the Luckiest</i>: the rare, good movie. It's very good.)</p> <p>User:口碑怎么样啊?(How is the reputation?)</p> <p>Bot:口碑还可以。(Good reputation.)</p> <p>User:那我要看看。(I want to see it.)</p> <p>Bot:你一定会喜欢的!(You will like it!)</p> <p>User:嗯,谢谢你,拜拜!(Well, thank you, bye!)</p> <p>Bot:拜拜!(Bye bye!)</p>	<p>User:今天几号了?(What's the date today?)</p> <p>Bot: 今天是2018年11月13日。(Today is Nov. 13, 2018.)</p> <p>User:你真棒。(You are so good.)</p> <p>Bot: 今天也是黄晓明生日呢,他是2018年出生。(Today is The birthday of Xiaoming Huang, who was born in 2018.)</p> <p>User:今天啊,我偶像啊。(Today! He is my idol.)</p> <p>Bot: 哈哈,那你喜欢看他的电影吗?我给你推荐一部他的电影《消失的子弹》,电影很不错,你可以去看看。(Haha, do you like his movies? I recommend you one of his movies, <i>The Bullet Vanishes</i>. It's a good movie. You can go check it out.)</p> <p>User:这不是他的电影(That is not his movie.)</p> <p>Bot: 希望你能喜欢! (I hope you like it!)</p> <p>User:不是很感兴趣,我以后再看看。(I'm not very interested, I will see later)</p> <p>Bot: 好的,再见。(OK, goodbye.)</p> <p>User:嗯,谢谢你,拜拜!(Well, thank you, bye!)</p> <p>Bot: 拜拜!(Bye bye!)</p>	<p>User:今天几号了?(What's the date today?)</p> <p>Bot: 今天是2018年1月13日。(Today is Jan. 13, 2018.)</p> <p>User:好的。(OK.)</p> <p>Bot: 你知道吗今天也是黄晓明生日呢!(Do you know, today is also Xiaoming Huang's birthday!)</p> <p>User:今天啊,我偶像啊。(Today! He is my idol.)</p> <p>Bot: 哈哈,那你知道吗,他也是第11届联合国艾滋病规划署中国亲善大使。(Haha, did you know that he is also the 11th UNAIDS Goodwill Ambassador for China!)</p> <p>User:是的,他还是第10届华语电影传媒大奖最具人气男演员呢!(Yes, he's also the most popular actor at the 10th Chinese Film Media Awards.)</p> <p>Bot: 对了,你知道吗,他有一首《撒娇女人最好命》,很不错的一首歌。(By the way, do you know, he has a song <i>Women Who Know How to Flirt Are the Luckiest</i>. Oh, a very good song.)</p> <p>User:这首歌好听吗?(Is this song good?)</p> <p>Bot: 这首歌不错,希望你能喜欢。(This song is good! I hope you like it.)</p> <p>User:嗯,谢谢你,拜拜!(Well, thank you, bye!)</p> <p>Bot: 拜拜!(Bye bye!)</p>
<p>User Profile</p> <p>Name: 胡振宇 (Zhenyu Hu)</p> <p>Gender: 男 (male)</p> <p>Favorite star: 黄晓明 (Xiaoming Huang)</p> <p>Ground-Truth Subgoals</p> <ol style="list-style-type: none"> 1. QA about the date 2. Chitchat about the star 3. Movie recommendation 	<p>Knowledge</p>	

Figure 5: Conversations produced by Trans.+kg, MGCG_G, and KERS. The red words indicate correct knowledge generated in the responses. The blue words are the usage of incorrect or inappropriate knowledge by models.

from the test scenarios. In total, 60 different scenarios are tested. After conversing with the dialog model, evaluators are asked to measure the dialog in terms of recommendation success, coherence, and engagingness with a 5-point Likert scale.

As shown in the right portion of Table 3, our model achieves a significant improvement in all the three metrics. It shows that KERS can complete different dialog types and finally make successful recommendations better than the baseline models.

6.3 Pair-wise Preference Test

We also conduct pair-wise comparisons on our model against baseline models. We ask ten evaluators to talk to both models under the same 60 scenarios selected in the dialog-level evaluation and select the better model. We show results in Table 4. KERS (t-test, $p < 0.05$) is preferred by evaluators over MGCG_G and Trans.+kg. This suggests KERS performs better than previous state-of-the-art models.

7 Case Study

To show the models' recommendation quality, we provide some examples. As shown in Table 5, KERS first answers the user's question correctly and talks about his favorite star *Xiaoming Huang* to engage the user. KERS then talks about *Xiaoming Huang's* awards and honors which gains user's trust.

Finally, KERS successfully recommends the movie *Women Who Know How to Flirt Are the Luckiest* starring *Xiaoming Huang* to users. Compared to KERS, MGCG_G recommends the inappropriate movie *The Bullet Vanishes* that is unrelated to the user's preferred star *Xiaoming Huang*. Trans.+kg recommends the correct movie title but mistakenly thinks *Women Who Know How to Flirt Are the Luckiest* is a song. We can also find that without the precise control of knowledge-aware response generation, both MGCG_G and Trans.+kg usually give wrong answers to questions. These observations indicate that accurate and rich knowledge is significant for the recommendation process.

8 Conclusions

It is vital to provide an informative and appropriate recommendation process in conversational recommendation with multiple dialog types. To improve recommendation quality, we present KERS to enhance the generated knowledge's accuracy and richness in responses. Our model uses a dialog guidance module to provide the proper subgoals and candidate knowledge, ensuring that the model interacts with the user in a planned way. In addition, we propose three new mechanisms: a sequential attention mechanism, a noise filter, and a knowledge enhancement module in the decoder. These mechanisms work together to increase the amount

and accuracy of knowledge in responses. Experimental results show that KERS completes various subgoals and obtains state-of-the-art results compared to previous models. In the future, we plan to further leverage knowledge graph’s path to enhance natural topic transitions in dialogs.

9 Ethical Considerations

Recently, recommendation dialog systems have developed rapidly, and we must consider ethical principles in both the design and development stages. First, The ultimate goal of the recommendation system is to provide users with content that they need. Therefore, the recommended content needs to be fair. The over-recommendation of a certain content due to the business relationship of interest undermines fairness. Second, the internal mechanism of the system must be transparent, so that users have a way to understand the nature of the system to avoid malicious sales. Similarly, during the operation of the recommendation dialog system, the collection of user information must be approved by the user to prevent the system from being used to collect user privacy. Finally, the recommended content cannot be factually false or misleading. For example, recommending misleading news will lead to the spread of rumors. The system needs to monitor the recommended content to solve such problems.

Acknowledgement

This research is funded by the Science and Technology Commission of Shanghai Municipality (20511101205), Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University (2020KEY001), and Xiaoi Research.

References

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.

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*, pages 139–148.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 815–824.

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.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxi-aoyang 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A survey on conversational recommender systems. *arXiv preprint arXiv:2004.00646*.

Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul A 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.

Yoon Kim. 2014. *Convolutional neural networks for sentence classification*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Sunhwan Lee, Robert Moore, Guang-Jie Ren, Raphael Arar, and Shun Jiang. 2018. Making personalized recommendation through conversation: Architecture design and recommendation methods. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2073–2083.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

- Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *Advances in neural information processing systems*, 31:9725–9735.
- 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.
- Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.
- Kevin Reschke, Adam Vogel, and Dan Jurafsky. 2013. Generating recommendation dialogs by extracting information from user reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 499–504.
- Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 235–244.
- Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.
- Zhuoran Wang, Hongliang Chen, Guanchun Wang, Hao Tian, Hua Wu, and Haifeng Wang. 2014. Policy learning for domain selection in an extensible multi-domain spoken dialogue system. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 57–67.
- Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2018. Transfertransfo: A transfer learning approach for neural network based conversational agents. In *NIPS2018 CAI Workshop*.
- Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.
- Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 177–186.
- 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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1006–1014.
- Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020b. Towards topic-guided conversational recommender system. *arXiv preprint arXiv:2010.04125*.