A Goal Interaction Graph Planning Framework for Conversational Recommendation

Xiaotong Zhang, Xuefang Jia, Han Liu*, Xinyue Liu, Xianchao Zhang

Dalian University of Technology, Dalian, China

zxt.dut@hotmail.com, jiaxuefang@hotmail.com, liu.han.dut@gmail.com, xyliu@dlut.edu.cn, xczhang@dlut.edu.cn

Abstract

Multi-goal conversational recommender system (MG-CRS) which is more in line with realistic scenarios has attracted a lot of attention. MG-CRS can dynamically capture the demands of users in conversation, continuously engage their interests, and make recommendations. The key of accomplishing these tasks is to plan a reasonable goal sequence which can naturally guide the user to accept the recommended goal. Previous works have demonstrated that mining the correlations of goals from the goal sequences in the dialogue corpus is helpful for recommending the goal that the user is interested in. However, they independently model correlations for each level of goal (i.e., goal type or entity) and neglect the order of goals appear in the dialogue. In this paper, we propose a goal interaction graph planning framework which constructs a directed heterogeneous graph to flexibly model the correlations between any level of goals and retain the order of goals. We design a goal interaction graph learning module to model the goal correlations and propagate goal representations via directed edges, then use an encoder and a dual-way fusion decoder to extract the most relevant information with the current goal from the conversation and domain knowledge, making the next-goal prediction fully exploit the prior goal correlations and user feedback. Finally we generate engaging responses based on the predicted goal sequence to complete the recommendation task. Experiments on two benchmark datasets show that our method achieves significant improvements in both the goal planning and response generation tasks.

Introduction

Conversational recommendation system (CRS) supports its users to achieve recommendation-related goals through a multi-turn dialogue (Jannach et al. 2022; Lei et al. 2020), and has been widely applied to voice assistants (Jannach et al. 2022), sociable recommendation (Hayati et al. 2020), and the e-commerce sites (Reschke, Vogel, and Jurafsky 2013; Fu et al. 2020), which not only improves user experience by providing them with rich choices, but also provides great commercial value (Jannach and Jugovac 2019; Chen et al. 2019). However, previous methods often assume that users clearly know their preferences, disregarding the fact that they may make the final decisions during their interactions with the system (He et al. 2017). Furthermore, most methods tend to focus solely on a single type of conversation, such as recommendation.

Recently, multi-goal conversational recommender system (MG-CRS) (Liu et al. 2020) is proposed, which contains various types of conversations, e.g. chitchat and Question-Answering (QA). MG-CRS aims to naturally guide the conversation from non-recommended scenarios to recommended scenarios, and proactively recommend to the user, which is more closer to realistic scenarios (Ram et al. 2018; Wang et al. 2014). Figure 1 shows a conversation in MG-CRS, along with the user profile, the domain knowledge, and a goal sequence with each goal containing two levels of goals (goal type and goal entity). The two-level goal sequences allow for greater control over the flow of dialogue and keep the dialogue as consistent as possible. Planning reasonable goal sequences for dialogues is a core issue in MG-CRS. Guided by the planned goal sequence, the system controls the transition of the dialogue and naturally leads the users to accept the goal at the end of the sequence.

To tackle this challenging task, MGCG (Liu et al. 2020) uses a CNN-based classifier to plan the goal sequences with dialogue history and the previous goal. Some methods such as TCP (Wang, Lin, and Li 2022), UniMIND (Deng et al. 2022), and TPNet (Wang, Lin, and Li 2023b) regard the goal planning task as a seq2seq paradigm (Sutskever, Vinyals, and Le 2014). In particular, TCP constructs a target-driven conversation planner to plan the goal path. UniMIND generates goals by finetuning a pretrained language model. TPNet is also a target-driven method and employs task-specific prompts for dialogue generation. Although these methods have considered various dialogue information as shown in Figure 1, but they neglect to capture the correlations between goals, which is helpful for predicting subsequent goals (Liu et al. 2022). To solve this issue, MGNN (Liu et al. 2022) constructs a hierarchical goal graph to capture the correlations of goals, and learns goal representations through GCN (Kipf and Welling 2017) for information propagation. However, MGNN separately propagates the representations of goal types and entities by GCN, and the information of goal entities can only be transferred to the goal type via a topic gate, which cannot capture the high-order relations between different levels of goals. Moreover, the edges of the graphs

*Corresponding author.

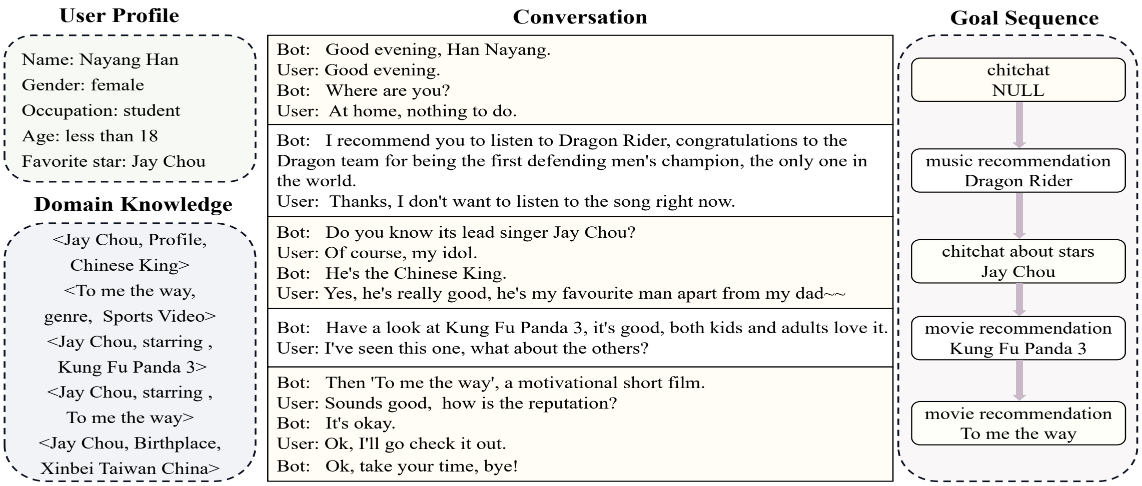

Figure 1: Illustration of the DuRecDial dataset. Each dialogue has four types of information, which are the user profile, the domain knowledge, the conversation itself, and the goal sequence. Each goal guides multiple turns of the dialogue, and contains a goal type and a goal entity which denote the dialogue type and the topic of the current turn in a dialogue, respectively. Both the goal type *movie recommendation* and the goal entity *To me the way* constitute the final goal that the user accepts.

are undirected, ignoring the order of goals appear in the dialogue.

In this paper, we propose a goal interaction graph planning framework to effectively plan goal sequences and use the planned goals to guide the response generation, which fully utilizes the correlation of goals and different types of input information in Figure 1. In particular, we first design a goal interaction graph learning module to model the goal relations with a directed heterogeneous graph based on domain knowledge and global goal sequences in dialogue corpus. Each node in the graph could be either goal type or entity, which enables the interactions between different levels of goals and implicitly models their high-order relations. Then we propagate the representations of goal types and entities with the heterogeneous graph attention neural network (Lv et al. 2021), which could dynamically assign weights to the edges. After that, an encoder module is proposed to integrate the different types of inputs into an informative representation. Then we propose a dual-way fusion decoder to extract the most relevant information with the current goal from the informative representation to complement the goal representations and facilitate the goal planning. Finally, the planned goal sequences serve as a guide for response generation to proactively reach the final goal that the user accepts. To the best of our knowledge, it is the first work that attempts to simultaneously model the goals at different levels using a heterogeneous graph. Experiments on two real-world datasets demonstrate the superiority of our approach in more proactively guiding the conversation, engaging the user, and achieving the final recommendation goal.

Related Work

Conversational recommendation system (CRS) can interact with users through a multi-turn dialogue and estimate reliable user preferences based on user responses (Wang and Benbasat 2013). Recently, goal-guided conversation system (Tang et al. 2019; Kishinami et al. 2022; Zhou et al. 2020a) has been proposed to better control the transition of dialogue, and each dialogue has a clear goal that proactively guides the flow of the dialogue. Kang et al. (Kang et al. 2019) collect a goal-driven recommendation dialogue dataset named GoRecDial, which introduces goals in the dialogue and leverages GoRecDial to implement end-to-end conversation recommendation. Wu et al. (Wu et al. 2019) create a goal-guided conversation dataset named Konv which contains about 270k utterances and 30k dialogues. In this challenging task, the model is trained to plan the goals on the knowledge graph to proactively lead the conversation. The explicit goal sequence planning work (Xu et al. 2020) further proposes a three-layer Knowledge-aware Hierarchical Reinforcement Learning based Model (KnowHRL) to improve user-interest consistency and dialogue coherence. Liu et al. (Liu et al. 2020) take a radical step by incorporating goal-guided conversations and knowledge-based conversations into CRS as multi-goal conversational recommender systems (MG-CRS). MG-CRS is closer to free-form CRS, but differs from it in that MG-CRS has more than one type of conversations such as chitchat and QA, which are more in line with realistic scenarios (Zhou et al. 2020b). The goals in MG-CRS are not limited to a given topic or focus on a single type of goal, but are a goal sequence of the conversation, which is a significant difference between MG-CRS and some early goal-guided conversations.

Currently, some studies are devoted to MG-CRS. MGCG (Liu et al. 2020) divides the goal planning task into two parts: goal completion estimation and current goal prediction. For goal prediction task, MGCG uses a CNN-based classifier (Kim 2014) commonly used for classification tasks. Then based on the planning goal, MGCG designs the retrieval-based model and generation-based model to generate responses respectively (Wu et al. 2019). TCP (Wang, Lin, and Li 2022) treats the goal planning task as a generative task, which generates a goal path containing both dialogue types and entities of the current turn. TPNet (Wang, Lin, and Li 2023b) has improved TCP by adding task-specific prompt to enhance generated dialogues. UniMIND (Deng et al. 2022) regards the goal planning task and the response generation task in MG-CRS as a unified framework, which uses a pretrained language model such as BART (Lewis et al. 2020) and T5 (Raffel et al. 2020) to generate responses. Furthermore, to enhance the capability of goal planning, a graph-grounded goal planning framework (MGNN) (Liu et al. 2022) is proposed to redesign the goal planning module with a hierarchical goal graph and learns goal representations via GCN. COLOR (Wang, Lin, and Li 2023a) utilizes the Brownian bridge stochastic process to predict the goal, and in particular, it divides the datasets into in-domain and out-of-domain test sets. Some works such as GOKC (Bai et al. 2021) and KERS (Zhang et al. 2021) assume that the goals of the dialogues already exist, which ignores the importance of goal planning module and focuses on the task of response generation. GOKC puts emphasis on generating responses that are consistent with knowledge, while KERS emphasizes knowledge selection and generates responses by a knowledge enhancement mechanism. KPN (Zhu et al. 2021) focuses on the selection of knowledge and has introduced a knowledge prediction module which aims to predict relevant knowledge and utilizes the predicted knowledge to help select the final answer.

The Proposed Method

Problem Formulation

Given a dialogue $d = \{\mathcal{H},\mathcal{U},\mathcal{K},\mathcal{G}\}$ in a dialogue corpus $\mathcal{D} = \{d_i\}_{i=1}^{N}$, where N is the total number of conversations in the dialogue corpus, $\mathcal{U} = \{u_i\}_{i=1}^{N_U}$ is the user profile associated with the user and N_U is the number of sentences in the user profile, $\mathcal{K} = \{k_i\}_{i=1}^{N_K}$ is the domain knowledge related to the dialogue and N_k is the number of knowledge triplets (<subject, relation, object>) in the domain knowledge, $\mathcal{H} = \{x_i\}_{i=1}^{n}$ is the dialogue context with a sequence of utterances, $\mathcal{G} = \{g_i\}_{i=1}^{n}$ is the goal sequence of $\mathcal{H}$, and n is the number of turns of the context $\mathcal{H}$. As mentioned before, the dialogue goal has two levels, which are defined as $g = \{g^t, g^e\}$, where g^t and g^e indicate the goal type and the goal entity in a turn. Our task is to plan the subsequent goal sequences $\{g_{n+1}, g_{n+2}, ..., g_{f-1}, g_f\}$, and generate the responses $Y = \{y_{n+1}, y_{n+2}, ..., y_f\}$ based on the planned goals to proactively guide the conversation to make the user accept the final goal g_f while engaging the user interest as much as possible.

Our model is divided into two modules: (1) *Goal planning module* plans an appropriate goal sequence to control the transition of the conversation. The module takes a dialogue $d = \{\mathcal{H},\mathcal{U},\mathcal{K},\mathcal{G}\}$ as input, and outputs the predicted goal $g_{n+1} = \{g^t_{n+1}, g^e_{n+1}\}$ of the next turn. (2) *Response generation module* generates a response based on the predicted goal g_{n+1} to lead the user to accept the recommendation.

Goal Planning

This component consists of three modules: *a goal interaction graph learning module*, *an encoder module* and *a dual-way fusion decoder module*. The framework of the goal planning module is shown in Figure 2.

Goal Interaction Graph Learning The construction of goal interaction graph relies on the domain knowledge and the global goal sequences in the dialogue corpus. Domain knowledge implies the correlations of different entities while the global dialogue goal sequences can capture the transitions and dependencies of the two levels of goals. In detail, for goals at the same level, the global dialogue goal sequences provide important prior information about the transition of goals. For goals at different levels, the global dialogue goal sequences contain the dependencies between the two levels of goals. The goal representations can be facilitated by interacting with each other through the goal heterogeneous graph.

Goal Interaction Graph Construction. The goal interaction graph is defined as $G = \{\mathcal{V}, \mathcal{E}\}$. $\mathcal{V}$ contains two kinds of nodes: goal types $\{g^t_i\}_{i=1}^{N_t}$ and goal entities $\{g^e_i\}_{i=1}^{N_e}$, where N_t and N_e are the number of goal types and entities, respectively. $\mathcal{E}$ contains three kinds of edges, which links the goal types, the goal entities, the goal types and entities, respectively. The goal interaction graph G is constructed based on three adjacency matrices $\boldsymbol{A}^t$, $\boldsymbol{A}^e$ and $\boldsymbol{A}^{te}$, in which an directed edge between two nodes is added if their adjacency value is greater than 0.

We first construct the *goal-type adjacency matrix* $\boldsymbol{A}^t$. The edges between two goal types are only derived from the global dialogue goal sequences since there are only correlations between goal entities in the knowledge graph. In particular, we capture the sequential co-occurrence frequencies of goals. In a dialogue goal sequence, if goal types g^t_i and g^t_j co-occur and appear sequentially next to each other, $[\boldsymbol{A}^t]_{ij}$ will be increased by one. After going through all the goal sequences in the dialogue corpus, we obtain the goal-type adjacency matrix $\boldsymbol{A}^t$, which involves the transition information of goal types.

Similar to $\boldsymbol{A}^t$, the *goal type-entity adjacency matrix* $\boldsymbol{A}^{te}$ is also calculated based on the global dialogue goal sequences. If a goal type g^t_i and a goal entity g^e_j are not only in the same conversation, but also co-guide the same conversation session, $[\boldsymbol{A}^{te}]_{ij}$ and $[\boldsymbol{A}^{te}]_{ji}$ will plus one.

The *goal-entity adjacency matrix* $\boldsymbol{A}^e$ is constructed based on both the global dialogue goal sequences and the domain knowledge. In particular, we first construct an adjacency matrix $\boldsymbol{S}$ which contains the sequential co-occurrence frequencies between two goal entities, which are as the same as the constructions of $\boldsymbol{A}^t$. Then we construct an adjacency matrix $\boldsymbol{K}$ which extracts the correlations between the goal entities from the domain knowledge, i.e., if the goal entity g^e_i and goal entity g^e_j are connected by a relation in the knowledge triplets, both $\boldsymbol{K}_{ij}$ and $\boldsymbol{K}_{ji}$ will plus one. $\boldsymbol{A}^e$ is the combination of $\boldsymbol{S}$ and $\boldsymbol{K}$, and defined as:

$$\boldsymbol{A}^e = \alpha\boldsymbol{S} + (1-\alpha)\boldsymbol{K}, \tag{1}$$

where α is the weight coefficient to balance the sequence

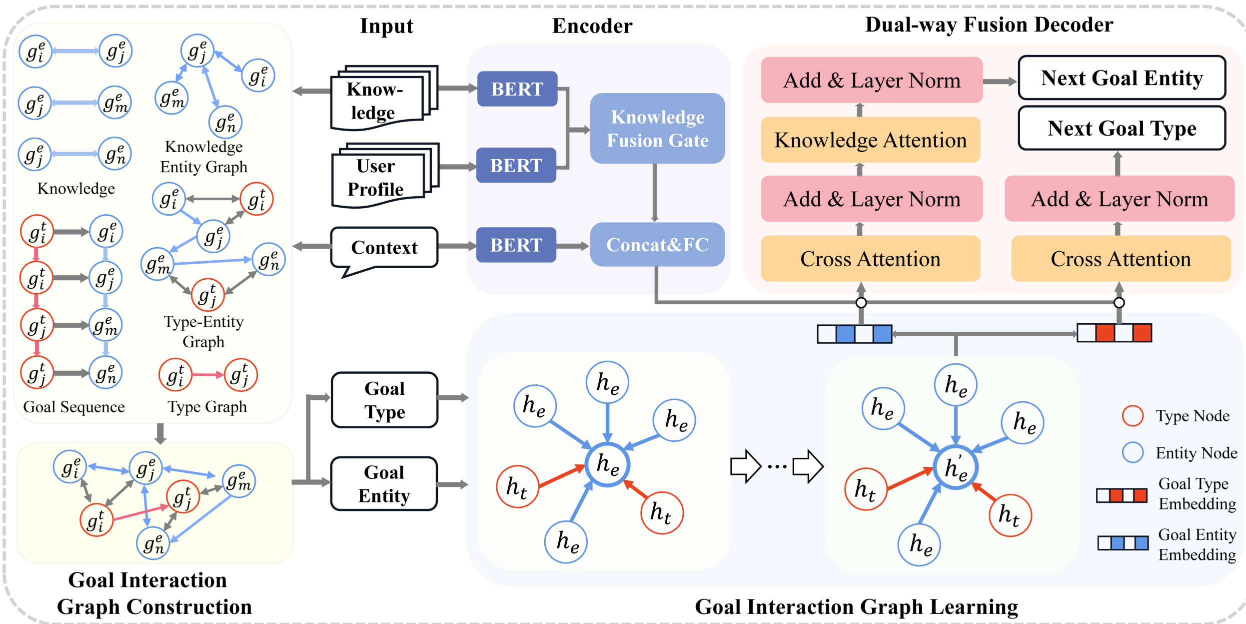

Figure 2: The framework of the goal planning module. We first construct a goal interaction graph by taking the goal types and entities in the dialogue corpus as nodes, and link the nodes based on the domain knowledge and the global goal sequences of all the dialogues. A heterogeneous graph attention neural network is built upon the goal interaction graph to learn the embeddings of goal types and entities. Then the domain knowledge, user profile and the context of conversation are encoded as conversation representations for goal planning. Finally a dual-way fusion decoder enhances the representations of the current goal type and entity with the conversation representation to predict the goal type and entity in the next turn of the dialogue.

information and the domain knowledge. By Eq. (1), two goal entities are highly correlated if they are not only sequentially next to each other in a conversation but also connected in the domain knowledge.

Goal Reprensentaion Learning. We learn the embeddings of the goals using the graph attention mechanism (Velickovic et al. 2018) through the goal interaction graph, which updates the representation of each goal by aggregating the information from its neighbouring nodes. Considering that the graph G is a heterogeneous graph with two types of nodes and three types of edges, we adopt a heterogeneous graph attention neural network, which is a variant of the graph attention network (Lv et al. 2021). For node g_i in the l-th layer of the network, the node embedding $\boldsymbol{h}_i^{(l)}$ is computed by quantifying the influence of the neighbouring nodes:

$$\boldsymbol{h}_i^{(l)} = \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^{(l)} \boldsymbol{W}_h^{(l)} \boldsymbol{h}_j^{(l-1)}\right), \quad (2)$$

where σ is LeakyReLU, α_{ij} is the attention score between the node g_i and its neighbouring node g_j, $\boldsymbol{W}_h^{(l)}$ is the weight matrix at the l-th layer. $\mathcal{N}_i$ is the set of neighbours of node g_i. Note that the neighbouring nodes could be goal types or goal entities, we distinguish them by different edge types. When updating the node representations in the graph, it is crucial to consider the information of edges and different types of adjacent nodes. Therefore, we incorporate them into the attention score $\alpha_{ij}^{(l)}$:

$$\alpha_{ij}^{(l)} = \frac{exp(\sigma(\boldsymbol{a}^T[\boldsymbol{W}_v\boldsymbol{h}_i||\boldsymbol{W}_v\boldsymbol{h}_j||\boldsymbol{W}_e\boldsymbol{e}_{ij}]))}{\sum_{k\in\mathcal{N}_i} exp(\sigma(\boldsymbol{a}^T[\boldsymbol{W}_v\boldsymbol{h}_i||\boldsymbol{W}_v\boldsymbol{h}_k||\boldsymbol{W}_e\boldsymbol{e}_{ik}]))}, \quad (3)$$

where $\boldsymbol{a}$ is a weight vector, $\boldsymbol{W}_v$ and $\boldsymbol{W}_e$ are the weight matrices for nodes and edges, respectively. $\boldsymbol{e}_{ij}$ is the edge embedding between the node g_i and node g_j.

After the goal interaction graph learning, the representations of goal types and entities are enhanced, since they not only integrate the global goal sequences from the dialogue corpus, but also learn the relevance between goals from the domain knowledge.

Encoder We design *a context encoder*, *a knowledge encoder* and *a user profile encoder* to embed different inputs related to each conversation, then combine them to learn the conversation representation for each dialogue.

For the context encoder, we directly use the BERT (Devlin et al. 2019) model to encode the dialogue context $\mathcal{H}$ and obtain the context representation for each conversation:

$$\boldsymbol{Enc}_H = \text{BERT}(x_1, x_2, ..., x_n), \quad (4)$$

where n is the length of dialogue history.

We treat both domain knowledge $\mathcal{K}$ and user profile $\mathcal{U}$ as external knowledge. Here we view the domain knowledge information as a whole rather than only considering the local relevance of the goal entities in the goal interaction graph. The BERT model is used to encode $\mathcal{K}$ and $\mathcal{U}$ separately, i.e.,

$$\boldsymbol{Enc}_K = \text{BERT}(k_1, k_2, ..., k_{N_K}), \quad (5)$$

$$\boldsymbol{Enc}_U = \text{BERT}(u_1, u_2, ..., u_{N_U}). \quad (6)$$

where N_K is the number of knowledge triples and N_U is the number of user profiles in the conversation, $\boldsymbol{Enc}_K$ focuses on the domain knowledge related to the conversation, whereas $\boldsymbol{Enc}_U$ focuses more on the user personal information. The two types of external knowledge representations are merged to produce a final external knowledge representation by a knowledge fusion gate, i.e.,

$$\boldsymbol{Enc}_K^{'} = \beta \boldsymbol{Enc}_K + (1-\beta)\boldsymbol{Enc}_U, \quad (7)$$

where β is a learnable weight to balance the domain knowledge and the user profile information. Since the current goal has a more important role for subsequent goal prediction, the calculation of β is based on the current goal, i.e.,

$$\beta = \text{GELU}(\boldsymbol{W}_\beta[\boldsymbol{h}_c^t || \boldsymbol{h}_c^e]), \quad (8)$$

where $\boldsymbol{h}_c^t$ is the embedding of current goal type and $\boldsymbol{h}_c^e$ is the embedding of current goal entity, $||$ indicates the concatenation operation, $\boldsymbol{W}_\beta$ is a weight matrix which is learnable in the training phase. After passing through the knowledge fusion gate in Figure 2, we concatenate the context representation and the external knowledge representation to obtain a high-quality conversation representation:

$$\boldsymbol{Enc} = \boldsymbol{W}_E(\boldsymbol{Enc}_K^{'} || \boldsymbol{Enc}_H), \quad (9)$$

where $\boldsymbol{W}_E$ is a weight matrix.

Dual-way Fusion Decoder The dual-way fusion decoder consists of two components: *the goal type decoder* and *the goal entity decoder*. For each dialogue, we integrate the embeddings of the goals and its conversation representation for predicting the goals in the next turn of the dialogue.

In the goal type decoder module, we first extract the most relevant information with the current goal type $\boldsymbol{h}_c^t$ from the conversation representation $\boldsymbol{Enc}$ with cross attention, and obtain the enhanced goal type representation:

$$\boldsymbol{O}_t = \text{Attention}(\boldsymbol{h}_c^t, \boldsymbol{Enc}, \boldsymbol{Enc}) + \boldsymbol{h}_c^t, \quad (10)$$

$$\text{Attention}(\boldsymbol{h}_c^t, \boldsymbol{Enc}, \boldsymbol{Enc}) = \text{softmax}(\frac{\boldsymbol{h}_c^t \boldsymbol{Enc}^T}{\sqrt{d_k}})\boldsymbol{Enc}, \quad (11)$$

where d_k is the dimension of goal embeddings. $\boldsymbol{O}_t$ incorporates the contextual and external knowledge information to facilitate the goal type prediction with a weight matrix $\boldsymbol{W}_{\hat{t}}$:

$$\hat{g}_{next}^t = \text{softmax}(\boldsymbol{W}_{\hat{t}} \boldsymbol{O}_t). \quad (12)$$

Similar to the goal type decoder, we first extract the most relevant information with current goal entity $\boldsymbol{h}_c^e$ from the conversation representation $\boldsymbol{Enc}$ with cross attention, and obtain the enhanced goal entity representation:

$$\boldsymbol{O}_e^{'} = \text{Attention}(\boldsymbol{h}_c^e, \boldsymbol{Enc}, \boldsymbol{Enc}) + \boldsymbol{h}_c^e. \quad (13)$$

Considering that domain knowledge representation contains many entity-related descriptions that may be useful for goal entity prediction, we further enhance the goal entity representation by involving domain knowledge information with knowledge attention, i.e.,

$$\boldsymbol{O}_e = \text{Attention}(\boldsymbol{O}_e^{'}, \boldsymbol{Enc}_K, \boldsymbol{Enc}_K) + \boldsymbol{O}_e^{'}. \quad (14)$$

The prediction of the goal entity in the next turn of the dialogue is computed with a weight matrix $\boldsymbol{W}_{\hat{e}}$:

$$\hat{g}_{next}^e = \text{softmax}(\boldsymbol{W}_{\hat{e}} \boldsymbol{O}_e). \quad (15)$$

Dataset	DuRecDial	DuRecDial 2.0
#Dialogues	10,190	6,292
#Utterances	155,477	99,869
#Domians	7	6
#Goal Types	21	19
#Goal Entities	701	646
#Knowledge Triples	222,198	123,298

Table 1: Dataset statistics.

Optimization The model is optimized with the cross-entropy loss between the true and the predicted goals in the training set. The whole goal planning loss is the sum of the goal type and goal entity losses:

$$L_G = -\frac{1}{N}\sum_{i=1}^{N} g_i^t \log \hat{g_i^t} - \frac{1}{N}\sum_{i=1}^{N} g_i^e \log \hat{g_i^e}, \quad (16)$$

where g_i^t, $\hat{g_i^t}$ and g_i^e, $\hat{g_i^e}$ are the true and predicted goal types and entities of the i-th sentence, respectively. Since we have learned representations for all the goal types and entities in the training phase, we only need to go through the encoder and the dual-way fusion decoder to predict the next goal in the inference phase.

Response Generation

To demonstrate the effectiveness of our goal planning model, we propose a simple yet effective response generation module which uses the planned goal to guide the response generation. Since the pretrained language model GPT-2 (Radford et al. 2019) has a good inference capability in generation task, we directly use GPT-2 as the backbone of our response generation module. The domain knowledge $\mathcal{K}$, the user profile $\mathcal{U}$, the conversation context $\mathcal{H}$, and the goals of the next turn in the dialogue are concatenated into a long sequence as the input of the model, where the goals are the concatenation of the goal type and the goal entity. Concretely, in the training phase, we take the true goal type and goal entity of the next turn as the input goals. Whereas in the inference phase, we use the goals predicted by the goal planning module as the input goals.

Experiment

We conduct experiments on both the goal planning task and the response generation task. The predicted goals in the goal planning module are used to guide the response generation.

Experimental Settings

Dataset We conduct experiments on two multi-type conversational recommendation datasets DuRecDial and DuRecDial 2.0 (Liu et al. 2020), which are commonly used in conversational recommendation. DuRecDial and DuRecDial 2.0 contain about 10k multi-turn Chinese dialogues and 6k multi-turn English dialogues, respectively. Each dialogue provides corresponding domain knowledge and user profiles. Following the previous works (Liu et al. 2020; Wang, Lin, and Li 2023b), we split the DuRecDial and DuRecDial 2.0 datasets into train/dev/test with 5,400/800/1,804 and 4,126/592/1,325 conversations, respectively. The statistics of these datasets are presented in Table 1.

Baselines To validate our method, we carefully select several competitive methods as our baselines in the goal planning task and the response generation task.

MGCG (Liu et al. 2020): It is a CNN based approach and uses a generation-based model and a retrieval-based model for response generation. We select the generation-based model as the baseline in the response generation task due to the poor performance of the retrieval-based model.

UniMIND (Deng et al. 2022): It is a prompt-based learning framework, which employs BART (Lewis et al. 2020) as the backbone.

TCP (Wang, Lin, and Li 2022): It is a target-driven dialogue generation model that treats goal planning as sequence generation task and uses GPT-2 for response generation.

MGNN (Liu et al. 2022): It is the state-of-the-art graph neural network based model for multi-goal recommendation systems, which is an improved version of MGCG model.

TPNet (Wang, Lin, and Li 2023b): It is a task-specific prompt-based approach similar to TCP, which uses a variety of pretrained models as the backbone, such as BART, DialoGPT and GPT-2. For the response generation task, we carefully choose BART (denoted as TPNet w/BART) and GPT-2 (denoted as TPNet w/GPT-2) as the baselines because BART and GPT-2 achieve better performance.

Implementation Details In our goal planning module, we set the embedding size to 768. The width of the hidden layer is as the same as the embedding size. We use Adam (Kingma and Ba 2015) to train the whole model. The learning rate and the weight decay rate are set to 2e-5 and 0.01. In the goal interaction graph learning module, we set the number of layers to 2 and the dimension of the edge embeddings to 64 with the dropout rate as 0.1. We set $\alpha = 0.42$ when constructing the goal-entity adjacency matrix A^e. In the response generation module, we use GPT-2_{base} (Wolf et al. 2019) as our backbone model. The embedding size is as the same as the goal planning module. We select the hyperparameters based on the validation set to get the optimal model. The experiments are performed in Python 3.7 and with NVIDIA GeForce RTX 3090.[1]

Evaluation Metrics To evaluate the goal planning module, we adopt Macro-averaged **Accuracy (Acc)** and **F1 score (F1)** that are widely used for classification task to avoid the class imbalance issue. To evaluate the response generation module, we follow the previous studies (Liu et al. 2020, 2022; Wang, Lin, and Li 2022) to evaluate the generated dialogues at the turn level of the conversation and use the metrics including the word-level **F1**, **BLEU** (Papineni et al. 2002), **Distinct (DIST-2)** (Li et al. 2016), **Knowledge F1** and **perplexity (PPL)** for evaluation. Specifically, **F1** and **BLEU** measure the word-level similarity and *n-gram* overlaps between the generated responses and the ground truth, respectively. **DIST-2** and **PPL** measure the diversity and the fluency of the generated responses, respectively. **Knowledge F1** measures the overlaps between the gold reference knowledge and the generated responses. We run the methods 10 times and report the average results with the best results highlighted in bold (GPT-2 with true goals are underlined).

[1]Our code is available at https://github.com/xfjdlut/GIGF.

Method	Goal Type		Goal Entity	
	Acc(%)	F1(%)	Acc(%)	F1(%)
MGCG	84.78	-	64.31	-
UniMIND	89.48	90.64	83.54	52.31
TCP	92.22	95.45	87.67	77.73
MGNN	95.46	96.87	88.32	80.69
TPNet	93.58	89.84	91.92	81.95
Ours	**97.41**	**97.88**	**93.47**	**82.75**
- (w/o Graph)	93.45	95.86	86.42	66.75
- (GAT)	94.10	96.19	88.17	73.02
- (w/o KnowledgeSelect)	93.95	96.33	89.31	77.84
- (w/o KnowledgeAttention)	93.51	95.96	87.49	73.06

Table 2: Results of goal planning on DuRecDial.

Method	Goal Type		Goal Entity	
	Acc(%)	F1(%)	Acc(%)	F1(%)
MGCG	85.23	-	57.68	-
UniMIND	96.51	87.86	86.55	72.82
TCP	91.32	83.68	87.04	74.14
MGNN	94.97	87.49	90.56	81.77
TPNet	93.21	85.79	83.26	78.59
Ours	**97.62**	**97.91**	**96.01**	**89.46**
- (w/o Graph)	96.05	96.14	91.53	81.72
- (GAT)	96.09	96.20	93.64	85.73
- (w/o KnowledgeSelect)	97.00	97.54	92.85	82.44
- (w/o KnowledgeAttention)	97.36	97.84	93.44	84.47

Table 3: Results of goal planning on DuRecDial 2.0.

Experimental Results

Results of Goal Planning The goal entity prediction is more difficult than the goal type prediction since the number of goal entities is much larger than the goal types. From the results in Table 2 and 3, it can be seen that our model achieves the best performance on the goal planning task, since it makes full use of the diverse inputs. In particular, our method not only constructs a directed heterogeneous graph to model various correlations between any level of goals to improve the goal representations, but also enhances the goal representations by extracting the most relevant information from different types of inputs. MGCG achieves poor results, because it simply uses two CNN classifiers to process all the input information. UniMIND, TCP and TPNet ignore to model the correlations between goals. It is worth noting that UniMIND has a topic prediction task and an item prediction task which are both for predicting goal entities. For the sake of fairness, we treat the topic and item prediction

Method	F1(%)	BLEU-1/BLEU-2	DIST-2	Knowledge F1(%)	PPL(↓)
MGCG	33.48	0.279/0.203	0.043	35.12	18.76
UniMIND	40.58	0.308/0.231	0.078	44.51	6.63
TCP	41.40	0.376/0.299	0.072	48.63	4.22
MGNN	43.50	0.382/0.274	0.064	45.00	6.47
TPNet w/BART	37.22	0.338/0.255	**0.083**	44.52	5.23
TPNet w/GPT-2	41.53	0.379/0.301	0.075	48.81	4.22
Ours	**47.52**	**0.426/0.348**	0.078	**56.02**	**3.04**
- (w/o Goal)	34.93	0.297/0.211	0.048	43.83	6.93
GPT-2 (true goals)	48.62	0.432/0.356	0.079	58.51	2.92

Table 4: Results of response generation on DuRecDial.

Method	F1(%)	BLEU-1/BLEU-2	DIST-2	Knowledge F1(%)	PPL(↓)
MGCG	32.26	0.293/0.182	0.051	29.35	26.68
UniMIND	33.66	0.238/0.184	**0.135**	23.75	6.81
TCP	33.12	0.295/0.201	0.070	34.86	5.44
MGNN	36.75	0.272/0.194	0.073	31.32	6.63
TPNet w/BART	36.28	0.296/0.204	0.093	40.22	6.33
TPNet w/GPT-2	34.62	0.308/0.217	0.082	38.80	5.87
Ours	**38.35**	**0.326/0.241**	0.089	**66.91**	**3.51**
- (w/o Goal)	31.58	0.282/0.175	0.049	25.88	6.95
GPT-2 (true goals)	38.52	0.326/0.241	0.089	66.96	3.51

Table 5: Results of response generation on DuRecDial 2.0.

as entity prediction and report the results. With the same response generation module as MGCG, MGNN prefers to model goal interactions via a hierarchical graph, which improves upon MGCG and the other baselines, and verifies the effectiveness of mining goal correlations in goal planning. Unfortunately, MGNN neglects to capture the high-order relations between different levels of goals and the order of goals through the graph, resulting in worse performance than our method. For running time, UniMIND takes 40 minutes per epoch on NVIDIA GeForce RTX 3090, while TCP and TPNet are almost as fast as our model, requiring approximately 1 hour per epoch on NVIDIA GeForce RTX 3090.

Results of Response Generation The results are shown in Table 4 and 5. Compared with MGCG that uses GRU based method for response generation, the other baselines generally have better generation performance due to the ability of the pretrained models for generating precise and fluent utterances. The performance gap between MGCG and MGNN demonstrates the importance of goal planning as their response generation modules have the same structure, but the goal planning performance of MGNN far exceeds that of MGCG. As the same as our model, the strong baselines TCP and TPNet also utilize GPT-2 for response generation. But our model performs much better than them, which demonstrates that the goals predicted in the goal planning module play a crucial role in generating responses.

Ablation Study

We conduct an ablation study for the goal planning module. The results are shown in the bottom of Table 2 and Table 3. We first remove the *goal interaction graph learning* module, and only use the encoder and the dual-way fusion decoder to predict goals (denoted as w/o Graph). The results show that the performance decreases sharply in all metrics, especially for the F1 score of the goal entity, which suggests that the goal interaction graph learning module plays a crucial role in goal planning. Considering that the performance of the goal entity degrades the most when the graph is removed, we construct the goal entity graph to further explore whether the goal interaction graph is valid. In particular, we construct a goal entity graph and perform graph attention network (denoted as GAT). It can be seen that the F1 score of the goal entity improves a little compared with the w/o Graph model, but still remains low since it only updates the representations of goal entities from the graph. This verifies that the interactions between goals can facilitate the representation learning of goals. We also remove the knowledge fusion gate in the *encoder* and the knowledge attention in the *dual-way fusion decoder*, which are denoted as w/o KnowledgeSelect and w/o KnowledgeAttention. It can be seen that their performance is lower than Ours, which indicates that external knowledge can provide important supplementary information in goal planning.

For the response generation module, we remove the predicted goals and use the true goals when generating responses, respectively, and denote them as w/o Goal and GPT-2 (true goals). The results in the bottom of Table 4 and Table 5 show that GPT-2 performs much worse without goal guidance and performs the best with the true goals. This verifies the importance of the planned goals of our model for response generation and recommendation.

User	Bot
Hi, Do you know who starred in the movie Really Love You?	**UniMIND: Starring Jacky Cheung.** **TCP: It's Jacky Cheung.** **Ours: Starring Jacky Cheung.**
Wow, you are really a smart guy. You know everything.	**UniMIND: Of course, I also know that he's a top ten favorite male star.** **TCP: Haha, I also know he is the best actor in Hong Kong Film Awards.** **Ours: Hey, I also know that he is the god of Chinese songs, and he won the Best Supporting Actor Award in the 27th Golden Horse Awards of Taiwan Film.**
Yeah, yeah, I know. He's my favorite!	**UniMIND: You have good taste. He won the Best Supporting Actor at the Hong Kong Film Award.** **TCP: He is one of the top ten outstanding young people in the world and the winner of the Golden Needle Award, the highest honor of the top ten Chinese golden songs.** **Ours: Yes, he won the best Actor in Hong Kong Film Awards.**
Well, he has won many awards, it is great, I really envy him!	**UniMIND: He starred in the The Days of Being Dumb is really interesting, recommend you to watch.** **TCP: And have you seen the A Chinese Ghost Story III in which he starred? The film's commercial design is too heavy, and it loses the first episode's human touch and sense of humor, but the special effects are still considerable.** **Ours: Why don't you take a look at him in A Chinese Ghost Story III? The film was too commercial and lost all the humanity and humour of the first episode, but there were still a lot of special effects.**
Thank you for your recommendation, but I prefer to listen to Leslie Cheung's music now.	**UniMIND: Well, listen to the thousands of songs he sings, Thousands of Songs (Live) is on your mind, every rhyme is told, and the parting scene is vividly displayed.** **TCP: I'd like to recommend one of his films, A Chinese Ghost Story III, there are still some amazing effects.** **Ours: Then you must not miss this dynamic Together Forever, Leslie Cheung once again shows his ability to control the song, in the state of the world without desire and desire, he is not so much singing a song, he is more like a band.**
User: It sounds very gentle.	**UniMIND: Yeah, it's really nice.** **TCP: Yeah, it's a great song.** **Ours: Would you like me to play this song for you?**
User: Yeah, I'd love to hear it. It must be great!	**UniMIND & TCP & Ours: Shall I play it for you now?**
User: Yeah, yeah, could you play it for me, please?	**UniMIND & TCP & Ours: Ok, it has been played for you, please enjoy.**
User: Yeah, I'm listening right now. Let's talk sometime. Bye!	**UniMIND & TCP & Ours: Okay, bye.**

Figure 3: Case study.

Case Study

To show the recommendation quality of our model, we provide some dialogues generated by our model and other baselines in Figure 3. The texts in red are the correct recommendations, whereas the texts in blue are wrong. In the first three turns of conversation, all the models respond correctly to the user. When it comes to recommend actor *Jacky Cheung's* movies to users in the fourth turn of conversation, our model and TCP are consistent with the ground-truth. Whereas UniMIND predicts the wrong goal *The Days of Being Dumb*, thus generating the response that deviates from the ground-truth response. When the user changes the subject and mentions that he now has a fondness for *Leslie Cheung's* music in the fifth turn, only our model accurately predicts the goal *Together Forever*. Based on the goal we predict, our model recommends *Leslie Cheung's* music to the user. Although UniMIND attempts to change the goal type to music recommendation, the song it suggests does not belong to *Leslie Cheung*. In contrast, TCP fails to switch the goal type thoroughly, and still tries to recommend movie *a Chinese Ghost Story* to user. Overall, our model is more coherent and informative when interacting with users.

Conclusion

In this paper, we first designed a goal interaction graph learning module, through which the representations of goals can be boosted with each other. Then we planned the goal path of the dialogue by incorporating contextual information and external knowledge into the goal representations via a dual-way fusion decoder. Finally we used the predicted goals to guide the response generation. The experiments show that our method achieves the state-of-the-art performance for both the goal planning and response generation tasks. In future, we plan to merge the goal planning and response generation modules into a unified model and incorporate the external knowledge more naturally into the responses.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62206038, 62106035, 61972065) and Fundamental Research Funds for the Central Universities (No. DUT20RC(3)066, DUT20RC(3)040) and National Key Research and Development Program of China (2022YFB4500300).

References

Bai, J.; Yang, Z.; Liang, X.; Wang, W.; and Li, Z. 2021. Learning to Copy Coherent Knowledge for Response Generation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 12535–12543.

Chen, Q.; Lin, J.; Zhang, Y.; Ding, M.; Cen, Y.; Yang, H.; and Tang, J. 2019. Towards Knowledge-Based Recommender Dialog System. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1803–1813.

Deng, Y.; Zhang, W.; Xu, W.; Lei, W.; Chua, T.-S.; and Lam, W. 2022. A Unified Multi-task Learning Framework for Multi-goal Conversational Recommender Systems. *ACM Transactions on Information Systems*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186.

Fu, Z.; Xian, Y.; Zhang, Y.; and Zhang, Y. 2020. Tutorial on Conversational Recommendation Systems. In *ACM Conference on Recommender Systems (RecSys)*, 751–753.

Hayati, S. A.; Kang, D.; Zhu, Q.; Shi, W.; and Yu, Z. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) (1)*, 8142–8152.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T. 2017. Neural Collaborative Filtering. In *World Wide Web (WWW)*, 173–182.

Jannach, D.; and Jugovac, M. 2019. Measuring the Business Value of Recommender Systems. *ACM Trans. Manag. Inf. Syst.*, 10(4): 16:1–16:23.

Jannach, D.; Manzoor, A.; Cai, W.; and Chen, L. 2022. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.*, 54(5): 105:1–105:36.

Kang, D.; Balakrishnan, A.; Shah, P.; Crook, P. A.; Boureau, Y.; and Weston, J. 2019. Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) (1)*, 1951–1961.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Kishinami, Y.; Akama, R.; Sato, S.; Tokuhisa, R.; Suzuki, J.; and Inui, K. 2022. Target-Guided Open-Domain Conversation Planning. In *International Conference on Computational Linguistics (COLING)*, 660–668.

Lei, W.; He, X.; de Rijke, M.; and Chua, T. 2020. Conversational Recommendation: Formulation, Methods, and Evaluation. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2425–2428.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 7871–7880.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 110–119.

Liu, Z.; Wang, H.; Niu, Z.; Wu, H.; Che, W.; and Liu, T. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 1036–1049.

Liu, Z.; Zhou, D.; Liu, H.; Wang, H.; Niu, Z.-Y.; Wu, H.; Che, W.; Liu, T.; and Xiong, H. 2022. Graph-Grounded Goal Planning for Conversational Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4923–4939.

Lv, Q.; Ding, M.; Liu, Q.; Chen, Y.; Feng, W.; He, S.; Zhou, C.; Jiang, J.; Dong, Y.; and Tang, J. 2021. Are we really making much progress?: Revisiting, benchmarking and refining heterogeneous graph neural networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 1150–1160.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Ram, A.; Prasad, R.; Khatri, C.; Venkatesh, A.; Gabriel, R.; Liu, Q.; Nunn, J.; Hedayatnia, B.; Cheng, M.; Nagar, A.; King, E.; Bland, K.; Wartick, A.; Pan, Y.; Song, H.; Jayadevan, S.; Hwang, G.; and Pettigrue, A. 2018. Conversational AI: The Science Behind the Alexa Prize. *CoRR*, abs/1801.03604.

Reschke, K.; Vogel, A.; and Jurafsky, D. 2013. Generating Recommendation Dialogs by Extracting Information from User Reviews. In *Annual Meeting of the Association for Computational Linguistics (ACL) (2)*, 499–504.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 3104–3112.

Tang, J.; Zhao, T.; Xiong, C.; Liang, X.; Xing, E. P.; and Hu, Z. 2019. Target-Guided Open-Domain Conversation. In *Annual Meeting of the Association for Computational Linguistics (ACL) (1)*, 5624–5634.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR) (Poster)*.

Wang, J.; Lin, D.; and Li, W. 2022. Follow Me: Conversation Planning for Target-driven Recommendation Dialogue Systems. *CoRR*, abs/2208.03516.

Wang, J.; Lin, D.; and Li, W. 2023a. Dialogue Planning via Brownian Bridge Stochastic Process for Goal-directed Proactive Dialogue. In *Findings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 370–387.

Wang, J.; Lin, D.; and Li, W. 2023b. A Target-Driven Planning Approach for Goal-Directed Dialog Systems. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, W.; and Benbasat, I. 2013. Research note¡ªa contingency approach to investigating the effects of user-system interaction modes of online decision aids. *Information Systems Research*, 24(3): 861–876.

Wang, Z.; Chen, H.; Wang, G.; Tian, H.; Wu, H.; and Wang, H. 2014. Policy Learning for Domain Selection in an Extensible Multi-domain Spoken Dialogue System. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 57–67.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR*, abs/1910.03771.

Wu, W.; Guo, Z.; Zhou, X.; Wu, H.; Zhang, X.; Lian, R.; and Wang, H. 2019. Proactive Human-Machine Conversation with Explicit Conversation Goal. In *Annual Meeting of the Association for Computational Linguistics (ACL) (1)*, 3794–3804.

Xu, J.; Wang, H.; Niu, Z.; Wu, H.; and Che, W. 2020. Knowledge Graph Grounded Goal Planning for Open-Domain Conversation Generation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 9338–9345.

Zhang, J.; Yang, Y.; Chen, C.; He, L.; and Yu, Z. 2021. KERS: A Knowledge-Enhanced Framework for Recommendation Dialog Systems with Multiple Subgoals. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) (Findings)*, 1092–1101.

Zhou, K.; Zhou, Y.; Zhao, W. X.; Wang, X.; and Wen, J.-R. 2020a. Towards Topic-Guided Conversational Recommender System. In *International Conference on Computational Linguistics (COLING)*, 4128–4139.

Zhou, L.; Gao, J.; Li, D.; and Shum, H. 2020b. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Comput. Linguistics*, 46(1): 53–93.

Zhu, Y.; Nie, J.; Zhou, K.; Du, P.; Jiang, H.; and Dou, Z. 2021. Proactive Retrieval-based Chatbots based on Relevant Knowledge and Goals. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2000–2004.