CRSAL: Conversational Recommender Systems with Adversarial Learning

XUHUI REN, HONGZHI YIN, and TONG CHEN, The University of Queensland, Australia
 HAO WANG, Alibaba Group, China
 NGUYEN QUOC VIET HUNG, Griffith University, Australia
 ZI HUANG, The University of Queensland, Australia
 XIANGLIANG ZHANG, King Abdullah University of Science and Technology, Saudi Arabia

Recommender systems have been attracting much attention from both academia and industry because of their ability to capture user interests and generate personalized item recommendations. As the life pace in contemporary society speeds up, traditional recommender systems are inevitably limited by their disconnected interaction styles and low adaptivity to users' evolving demands. Consequently, conversational recommender systems emerge as a prospective research area, where an intelligent dialogue agent is integrated with a recommender system. Conversational recommender systems possess the ability to accurately understand end-users' intent or request and generate human-like dialogue responses when performing recommendations. However, existing conversational recommender systems only allow the systems to ask users for more preference information, while users' further questions and concerns about the recommended items (e.g., enquiring the location of a recommended restaurant) can hardly be addressed. Though the recent task-oriented dialogue systems allow for two-way communications, they are not easy to train because of their high dependence on human guidance in terms of user intent recognition and system response generation. Hence, to enable two-way human-machine communications and tackle the challenges brought by manually crafted rules, we propose Conversational Recommender System with Adversarial Learning (CRSAL), a novel end-to-end system to tackle the task of conversational recommendation. In CRSAL, we innovatively design a fully statistical dialogue state tracker coupled with a neural policy agent to precisely capture each user's intent from limited dialogue data and generate conversational recommendation actions. We further develop an adversarial Actor-Critic reinforcement learning approach to adaptively refine the quality of generated system actions, thus ensuring coherent human-like dialogue responses. Extensive experiments on two benchmark datasets fully demonstrate the superiority of CRSAL on conversational recommendation tasks.

CCS Concepts: • **Information systems** → **Users and interactive retrieval**; **Personalization**; • **Human-centered computing**; → **Systems and tools for interaction design**;

Corresponding author; contributing equally with the first author.

The work has been supported by ARC Discovery Project (Grant Nos. DP190101985 and DP170103954) and National Natural Science Foundation of China (Grant No. 61672501).

Authors' addresses: X. Ren, H. Yin, and T. Chen, The University of Queensland, Australia; emails: xuhui.ren@uq.net.au, {h.yin1, tong.chen}@uq.edu.au; H. Wang, Alibaba Group, China; email: cashenry@126.com; N. Q. V. Hung, Griffith University, Australia; email: quocviethung1@gmail.com; Z. Huang, The University of Queensland, Australia; email: huang@itee.uq.edu.au; X. Zhang, King Abdullah University of Science and Technology, Saudi Arabia; email: xiangliang.zhang@kaust.edu.sa.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1046-8188/2020/06-ART34 \$15.00

<https://doi.org/10.1145/3394592>

Additional Key Words and Phrases: Conversational recommender systems, dialogue systems, adversarial learning, deep neural networks

ACM Reference format:

Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Nguyen Quoc Viet Hung, Zi Huang, and Xiangliang Zhang. 2020. CRSAL: Conversational Recommender Systems with Adversarial Learning. *ACM Trans. Inf. Syst.* 38, 4, Article 34 (June 2020), 40 pages.
<https://doi.org/10.1145/3394592>

1 INTRODUCTION

For e-commerce platforms, appropriate recommendations will offer users easier access to preferred resources and also help service providers understand their customers better. As such, the recommender system (RS) serves a pivotal purpose of aligning available items (e.g., products, services) with user interests. The recommender system, at its core, extracts user preference information from each user's behavior history and available profile and then recommends items that can best fit this user's preference [49].

Traditionally, recommender systems are designed as single-shot models [59] based on the information acquired from user ratings [63, 85] or interaction records [9, 26, 30, 92, 93]. Despite the tremendous amount of research attention, almost none of existing recommender systems can behave in a similar way to humans, rendering it difficult to capture users' dynamic intentions and demands when asked for a recommendation. Furthermore, the cold-start scenario in recommendation [66, 67] sets a tremendous barrier for traditional recommendation paradigms, where recommendations are needed to be generated for new users without any user-related prior knowledge. In contrast, in the offline real-life cases, when a new customer walks into a store, the sales manager would start a conversation with the customer to first identify the demand of the customer (e.g., buying a laptop), then narrow down the options according to customer's criteria (e.g., lightweight), and last present items that will most likely to satisfy the customer's needs (e.g., *Microsoft Surface*). Whenever this customer wants to revise or update her/his demand and preference (e.g., high performance), a short conversation can precisely deliver such information to the salesperson, and new recommendations will be given instantly (e.g., *MacBook Pro*). As shown in this real-world example, the capability of carrying out conversations is the key to unlock the limitations of existing recommender systems, as dialogue is the most natural and approachable way for end-users to express their intentions and demands. In short, compared with traditional recommender systems, a conversational recommender system (CRS) that "chats" with users is more appealing and beneficial to improve the recommendation performance and user experience.

To this end, we aim to advance the performance of recommender systems with the power of conversations. On one hand, the passive human-machine interaction form of traditional recommender systems poses a bottleneck on both user experience and model accuracy. On the other hand, the emerging research on the human-machine dialogue system (DS) [58, 79, 82] creates a bootstrap for our development of a conversational recommender system. Dialogue systems grant machines the capability to understand end-users' languages and imitate humans to generate human-like responses. Successful applications of DS, such as Apple Siri, Google Home, and Microsoft Cortina, have established the channel for human-machine conversations and advanced our daily lives with exceptional convenience. The existing researches about DS mainly focus on only one of the four functions [19]: chit-chat [86, 90], information search [16], task completion [3, 45], and document understanding [88]. DS for the general recommendation task has been rarely studied, and it is more complicated. In the real world, there are many scenarios where recommendations can only be presented via dialogues (e.g., a user is asking for a fuel station recommendation when driving)

Fig. 1. An example of recommending a restaurant to the customer through conversation.

instead of touch devices, conversational recommender systems are more advantageous in terms of human-machine interactions. However, as demonstrated in Figure 1, apart from finishing the task of recommending restaurants to a user, the conversational recommender system is also responsible for simple chit-chat (e.g., welcome greetings), information retrieval (e.g., ranking restaurants for each user), and understanding dialogue information (e.g., extracting user preferences from dialogues). As a result, a robust conversational recommender system requires much more than straightforwardly stacking a dialogue component upon a recommender system, and it is a non-trivial task to fuse two modules into a unified, end-to-end framework that delivers human-like conversations and make high-quality recommendations at the same time.

Besides, to make the conversational recommender system fully human-like, the challenges also arise from the practicality aspect. From the customers' perspective, through multi-turn conversations, the system should help the customers clarify their demands and find ideal items in an accessible way. Meanwhile, the conversational recommender system could answer users' further questions about the recommendations, such as the address of a recommended restaurant and the warranty policy of a laptop. When being released from manually searching for the answers to their concerns, users will be able to focus on the decision-making process.

The emerging research area of conversational recommender system has attracted many attempts in the last few years [37, 39, 58], and companies such as Microsoft, Google, and Alibaba have

Fig. 2. The architecture of our proposed CRSAL.

also started to get involved [7, 19, 40]. However, existing conversational recommender systems are mainly based on the “systems ask, users answer” paradigm [72, 95], which means only the conversational recommender systems can propose questions (e.g., preferred restaurant location) to acquire users’ demand (e.g., near the city) in several conversation rounds. However, this one-way communication scheme prohibits users from asking follow-up questions about the recommended item (e.g., does the restaurant offer carparks), rendering it difficult for such systems to fully address users’ inquiries and concerns. At the same time, some researchers view the conversational recommendation problem from the perspective of DS-based information retrieval and propose task-oriented dialogue systems [44, 82, 97] to achieve similar goals while allowing for two-way communications. However, these approaches suffer from the typical drawbacks of the current task-oriented dialogue systems. First, they require abundant human labor and domain knowledge to annotate the training corpus by constructing a comprehensive dictionary that links alternative notations of ontology items that vary lexically or morphologically (e.g., “expensive” and “luxury” express the same meaning as “high price” in the dictionary) [82]. Besides, these task-oriented dialogue systems need to train a policy agent, which chooses the most suitable dialogue action (e.g., ask questions, provide recommendations, give information) from a manually crafted explicit dialogue action pool to guide the generation of machine conversations [11, 70] in each conversation turn. Consequently, in task-oriented dialogue systems, both the diversity and accuracy of dialogue responses are hugely limited by these hand-crafted rules. Furthermore, recent task-oriented dialogue systems are fine-tuned with reinforcement learning (RL) by asking the end-users for an evaluation score as the reward at the end of a conversation [17, 45, 84], which is impractical in real-life application scenarios.

To tackle the aforementioned challenges, in this article, we propose a novel end-to-end Conversational Recommender System with Adversarial Learning (CRSAL) that integrates the dialogue functionality with recommender systems seamlessly. As shown in Figure 2, CRSAL consists of three key modules. First, the *information extraction* module analyzes the input user utterance to track the current dialogue state, extract the user’s current intent, and generate a personalized recommendation list. In this module, to avoid the disadvantages of pre-defined, inflexible semantic mapping dictionaries, we propose a fully statistical neural dialogue belief tracker, which utilizes the implicit, latent semantic representations of tokens to identify the user demand instead of performing explicit token matching. Afterwards, the outputs of the information extraction process, including the dialogue belief state, the utterance intent, and the personalized recommendation list, will be fed into the *conversational response generation* module. A neural policy agent is introduced here to generate dialogue actions, which will lead the natural language generator to decode

the information passed from the first module into conversations. Both the neural policy agent and natural language generator will be trained via a variational inference network. By adaptively determining the dialogue actions based on the semantically clustered utterance tokens, we bypass the limitations of manually designed dialogue actions, thus substantially improve the dialogue action generation performance during conversational recommendations. At last, to ensure the accuracy of generated dialogue actions, the *adversarial reinforcement learning* module is introduced to fine-tune the dialogue policy agent. We propose an innovative reinforcement learning method by extending the advantage Actor-Critic algorithm [22] to an adversarial learning scheme, which encourages the dialogue policy agent to fully mimic human-made responses (i.e., ground-truth). To better evaluate the effectiveness of the conversational recommender systems, we also propose a joint evaluation metric, namely, **QoCR**, which quantifies the overall performance by considering both the task success rate and the quality of system-generated responses.

In summary, this article has made the following contributions:

- We propose CRSAL, a novel conversational recommender system that can interact with end-users using human-like conversations, analyze user demands, and provide personalized conversational recommendations. All the components of CRSAL are modularly connected, enabling the system to be trained in an end-to-end fashion.
- In CRSAL, we establish advantages against existing conversational recommender systems with two-way human-machine communications, while overcoming the drawbacks of hand-crafting semantic dictionaries and dialogue responses with a fully statistical dialogue belief tracker and an innovative conversational response generation scheme based on neural variational inference networks.
- To make the generated dialogue responses more accurate and human-like, we further propose an adversarial training approach for CRSAL based on advantage Actor-Critic reinforcement learning algorithm to actively refine the quality of actions generated by the policy agent.
- We conduct extensive experiments on three public datasets, and the results show the superiority of CRSAL against state-of-the-art baselines in terms of both the recommendation effectiveness and conversation quality. A new joint evaluation metric is also proposed to evaluate the all-round performance of conversational recommender systems in real-life scenarios.

2 RELATED WORK

There are three lines of research related to our work, namely, recommender systems, task-oriented dialogue systems, and reinforcement learning. We will review these research areas in this section.

2.1 Recommender Systems

Recommender systems have become a significant component for online service providers in the era of e-commerce, because they are efficient tools for exploring user preferences and generating personalized recommendations. Many e-commerce companies such as Amazon and Netflix deploy recommender systems to analyze user preferences based on explicit and/or implicit feedback and infer their potentially preferred items [29, 74]. Based on the type of input data, the recommender systems can be summarized into collaborative filtering (CF) systems, content-based recommender systems, and hybrid recommender systems [10, 75, 94]. Early researches in CF formulate the recommendation task as a regression problem to predict the user potential rating score for the candidate items [36, 61]. However, the rating-based recommendation models cannot perform well in top-N recommendation, which arouses the attention on ranking-based recommendation [15]. The core idea of ranking-based recommendation is to learn a model based on the relative preference

of a user over pairs of items. Then the model can generate a pairwise user-item affinity score for personalized item ranking [31].

Recent years have witnessed great progress in ranking-based recommendation [12, 41, 76, 91]. For example, Rendle et al. first proposed Bayesian Personalized Ranking (BPR) [62] based on CF. Hsieh et al. discussed the limitation of feature interactions in Factorization Machines (i.e., inner product) and proposed the Collaborative Metric Learning (CML) framework to model feature interactions based on Euclidean distance of the feature vector [29]. More recently, He et al. combined the strength of FM and Multi-Layer Perceptions (MLPs) and introduced Neural Factorization Machines [23]. This model enhances the overall recommendation performance by learning the high-order and non-linear feature interactions. In addition, temporal patterns within user transactions have also motivated extensive research on session-based recommendation [21, 47].

However, most state-of-the-art recommender systems highly rely on users' historical trajectories to generate personalized recommendation [26]. It lacks the capability of capturing users' dynamic intentions/demands and recommending the item in a human-like way. In contrast, our proposed CRSAL integrates the dialogue system with a recommender system seamlessly to perform human-like conversational recommendations. The dialogue system generates suitable conversational responses to acquire dynamic user demands and present the recommended items through dialogues. Conversational recommender systems can behave like a human assistant to communicate with the user, which could greatly improve the using experience.

2.2 Task-oriented Dialogue Systems

Task-oriented dialogue systems are one important branch in dialogue system research, which aims to help the user finish some specific tasks [96]. Though conversational recommendation is an emerging research topic, the basic concepts and model designs are mostly originated from task-oriented dialogue systems. For example, Sun et al. [72] designed a conversational recommender system to process the semi-structured user queries with facet-value pairs and fine-tune the generated dialogue policies with deep reinforcement learning. Zhang et al. [95] proposed the "systems ask, users respond" (SAUR) paradigm for conversational recommendation in e-commerce. In SAUR, a system agent is designed to ask users different questions to obtain clarified demands continuously, and a multi-memory network is utilized to analyze the user utterances for recommendation. In the most recent, Li et al. [40] introduced a new neural architecture for conversational recommendation and constructed a movie dataset for evaluation, and Christakopoulou et al. built a conversational recommender system based on sequential neural networks [13]. However, existing conversational recommender systems simply regard human-machine interactions as an approach to let users detail their demands. As a result, these systems lack the capability of answering questions from users, which is actually important to fully address users' concerns about the suggested items and ensure successful recommendations.

Along another line, research on traditional task-oriented dialogue systems has achieved significant improvements in helping users finish some specific tasks via human-like dialogue responses. One of the earliest end-to-end task-oriented dialogue systems was proposed in Reference [97], which applied a deep recurrent neural network to a relatively simple "yes/no" question-answering task. More task-oriented dialogue systems are developed for various real-life tasks since then, such as searching for restaurants [82], ordering flight tickets [79], and booking for movies [58]. Wei et al. [79] presented a sequence-to-sequence dialogue system for flight ticket booking, which can generate high-level human-like responses. Unfortunately, the major limitation of task-oriented dialogue systems comes from their high dependency on the manually crafted rules. Existing task-oriented dialogue systems count on the predefined semantic matching dictionaries to identify the user demand and dialogue actions, which require huge domain knowledge and human labor. Moreover,

state-of-the-art approaches rely on reinforcement learning (RL) to fine-tune the dialogue policy agent. Such systems require end-users' feedback (e.g., ratings or binary evaluations) as the RL reward signal to fine-tune the model parameters. However, the feedback from the user is usually subjective and difficult to obtain in practice, making these models underperform in real applications.

The two-way interaction we emphasize in task-oriented dialogue systems has some similarities to the mixed initiative interaction in the information retrieval literature [1, 4, 14, 27]. On one hand, task-oriented dialogue systems are designed to help users complete some specific tasks through natural language response [87]—they mostly make machines ask users a series of questions to gather sufficient information to complete the task (e.g., extracting user preferences for recommendation). In CRSAL, we aim to achieve the “two-way” communication where both the user and machines can either ask questions or provide requested information. On the other hand, mixed initiative agents provide a flexible interaction scheme where the user and the agent can take turns to exchange initiatives [1, 4, 14, 27], but existing mixed initiative agents lack the ability to infer users' implicit personal preferences to retrieve the most suitable items from diverse candidate items for recommendation. Hence, our proposed system could be regarded as a recommendation-oriented dialogue system that enables flexible two-way interactions as in mixed initiative agents and is also able to model users' latent interests for personalized recommendation.

It is also worth mentioning that we assimilate inspirations from some recent research [20, 55, 89, 98] to solve the aforementioned problems. Our proposed CRSAL identifies the user demand using the semantic representation of tokens instead of explicit semantic matching and clusters the dialogue actions straightly from the raw data. In CRSAL, an adversarial advantage Actor-Critic RL approach is introduced to imitate real end-users to generate the evaluation feedback. In this way, we can eliminate the limitation of manually crafted rules and in turn improve the overall recommendation performance.

2.3 Reinforcement Learning

Exploiting reinforcement learning (RL) on the policy agents in task-oriented dialogue systems has recently demonstrated promising effectiveness [19, 58, 71, 72, 98]. RL-based dialogue systems assume the model will receive a reward signal at the end of a conversation in the form of either a continuous score or binary evaluation, thus helping optimize the model parameters. In practice, the feedback from real end-users is subject to biases and unavailability [70], which creates a significant obstacle for fully optimizing these dialogue systems.

Two major approaches have been proposed to tackle the aforementioned problem. One is to utilize prior knowledge learned from real conversations among humans coupled with imitation learning or rule-based learning methods [43, 68]. The dialogue agent first pre-trains an initial rule-based policy agent and then refines the policy agent with RL. The other is to introduce a heuristic reward function to guide the policy decision. For example, Su et al. proposed a reward estimation method through online learning with Gaussian process classification [69]. Houthoofd et al. studied an exploration strategy based on the maximization of information gain [28] by designing an intrinsic reward bonus to the reward function.

Meanwhile, the Actor-Critic-based (AC) RL method can combine the advantage of the above two approaches and achieve superior performance on optimizing the policy agent [80]. However, conventional AC methods suffer from the problem of high variance of gradients, which would result in an unstable learning process and/or sub-optimal policy agents [78]. Mnih et al. introduced a method named advantage Actor-Critic (A2C) reinforcement learning to solve the high variance problem by adding an advantage function [51]. This function could decrease the cumulative variance in reward signal and then achieve more stable updates. But the A2C approach still relies on the

prior knowledge to guide the learning process. Recent approaches of generative adversarial networks (GANs) have inspired much research in task-oriented dialogue systems [38, 44, 57]. Liu et al. proposed a sample method to use the adversarial learning to fine-tune the dialogue system [44]. Peng et al. combined the advantage Actor-Critic reinforcement learning with adversarial learning and proposed an adversarial reinforcement learning method [57]. This method stacks an traditional A2C reinforcement learning with an adversarial learning layer to formulate a two-phase reinforcement learning. This model is, however, hard to find a group of satisfying parameters, because he directly searches the optimized parameters with reinforcement learning on a random state and conduct twice parameters with two different gradients in one batch. The two-phase reinforcement learning also increases the computational cost during the training. Based on the previous work, we propose a novel adversarial advantage Actor-Critic (A3C) reinforcement learning method in this article. We simplified the previous adversarial advantage Actor-Critic reinforcement learning structure and embedded the adversarial learning directly into the reinforcement learning. We leverage the raw data to learn the policy agent (i.e., the Actor) and the discriminator. The discriminator is trained to discriminate the generated action from the human expert and dialogue policy agent as accurately as possible. And the output of the discriminator serves as an important intrinsic reward to construct the feedback from the “Critic,” which is sequentially used to encourage the Actor to generate human-like responses.

3 THE PROPOSED MODEL

In this section, we detail our proposed model, namely, Conversational Recommender System with Adversarial Learning (CRSAL). As illustrated in Figure 2, the workflow of CRSAL consists of three phases: information extraction, conversational response generation, and reinforcement learning. With a user utterance as the input, we perform a three-way information extraction with: (1) a dialogue state tracker (DST) to infer the current dialogue belief state from the utterance; (2) a neural intent encoder (NIE) to encode the utterance into a vectorized representation of the user’s intent; and (3) a neural recommender network (NRN) to derive personalized recommendations based on the dialogue states and item features. Then, based on the extracted information, the conversational response generation is accomplished by two components: (1) a neural policy agent (NPA) to determine the type of next system action to take conditioned on the current dialogue state; and (2) a natural language generator (NLG) to compose conversational responses regarding the selected system action. These two components are designed synchronously, and the related parameters are updated by a neural variational inference network. At last, in the reinforcement learning stage, an adversarial learning scheme based on advantage Actor-Critic algorithm is proposed to fully fine-tune the model.

3.1 Information Extraction

In CRSAL, the information extraction module is responsible for distilling semantic and preference information from raw user utterances. In what follows, we will introduce its three components, namely, the dialogue state tracker (DST), the neural intent encoder (NIE), and the neural recommender network (NRN).

3.1.1 Dialogue State Tracker. The dialogue belief state refers to the representation of each current linguistic semantics expressed via user utterances, and tracking dialogue states is the key to ensuring robustness in dialogue systems [6]. The dialogue state tracker (DST) is essentially responsible for extracting the latent linguistic features from the raw input texts. In a conversational recommender system, dynamically tracking dialogue belief states is the key for understanding users’ demands and generating coherent, context-sensitive responses [32]. Conventional methods

Fig. 3. The computational process of the fully statistical dialogue state tracker. Different information layers are differentiated by colors.

for dialogue state tracking mainly employ recurrent neural networks (RNN) and convolutional neural networks (CNN) [53] to extract n -gram features from delexicalized user utterances. Usually, the item properties (e.g., price) and corresponding adjectives (e.g., cheap) are, respectively, referred to as *slots* and *values* in dialogue systems. As an important preprocessing step, the delexicalization of an utterance replaces the item properties and adjectives mentioned by the user with generic symbols to enable easier modelling and eliminate ambiguity, e.g., “want a restaurant of cheap price” and “want a restaurant of low price” will both be delexicalized into “want a restaurant of [Value – Cheap] [Slot – Price Range]” before being used to extract semantic features.

However, as discussed in Reference [54], the delexicalization-based methods usually require handcrafted matching dictionaries or manually annotated mentions of ontologies, which can be very labor-intensive and time-consuming. Also, as different types of online service platforms come with specialized slots and values (e.g., the keywords for “food type” can be highly different on two restaurant recommendation websites), existing dialogue state-tracking methods are highly restricted in applicability and generalizability. To bypass the shortcomings of delexicalization-based state tracking methods, we propose a fully statistical DST based on deep neural networks. In our proposed DST, tokens are discriminated via learned word embeddings instead of explicit vocabulary matching schemes, which offers better state tracking effectiveness and flexibility.

Figure 3 presents the computational process of our proposed DST. At each conversation turn t , the DST treats the new user utterance input and updates the dialogue belief state $\mathbf{b}_t$, which is a multinomial distribution over each candidate slot-value pairs $\omega = (s, v) \in \Omega$. The slot-value pair $\omega = (s, v)$ is a tuple of the slot s (e.g., $s \in \{\text{price range, food type}\}$ for restaurants) and the corresponding slot value v (e.g., $v \in \{\text{high, medium, low}\}$ when $s = \text{price range}$). Each slot-value pair represents a user’s expectation on a particular property of the items to be recommended. In the DST, to model the semantic meanings within different slot-value pairs, we derive an embedding vector $\mathbf{c}_\omega$ for each specific slot-value pair ω . Note that in this article, we omit the computation process of all look-up embeddings to be succinct.

First, in the t th conversation round, given a user utterance $\mathcal{U}_t = \{\mathbf{e}_m^{tok}\}_{m=1}^M$, which is a sequence of M tokens represented by pre-trained word embeddings $\mathbf{e}$, we compute a synergic feature representation $\mathbf{u}_t$ of utterance $\mathcal{U}_t$. The overview of this computation is shown in Figure 4. To start with, we construct the n -gram feature matrix $\mathbf{G}_t^n \in \mathbb{R}^{n \times (M-n+1)}$ from $\mathcal{U}_t$:

$$\mathbf{G}_t^n = \left[\left[\mathbf{e}_1^{tok} \oplus \dots \oplus \mathbf{e}_n^{tok} \right] \left[\mathbf{e}_2^{tok} \oplus \dots \oplus \mathbf{e}_{n+1}^{tok} \right] \dots \left[\mathbf{e}_{k-n+1}^{tok} \oplus \dots \oplus \mathbf{e}_M^{tok} \right] \right], \quad (1)$$

where $\oplus$ is the vector concatenation operation. Following Reference [54], we leverage the unigram, bigram, and trigram features by setting $n = 1, 2, 3$ and construct three feature matrices $\mathbf{G}_t^1, \mathbf{G}_t^2, \mathbf{G}_t^3$. The synergic feature representation $\mathbf{u}_t$ is calculated as follows:

$$\mathbf{u}_t = \text{CNN}_1(\mathbf{G}_t^1) + \text{CNN}_2(\mathbf{G}_t^2) + \text{CNN}_3(\mathbf{G}_t^3), \quad (2)$$

Fig. 4. The CNN-based synergic representation learning. The convolutional filters are adopted to n-gram utterance features (e.g., unigram feature G^1 , bigram feature G^2 , trigram feature G^3). The final representation u is the summation of all three intermediate features.

where $CNN_1(\cdot)$, $CNN_2(\cdot)$, $CNN_3(\cdot)$ denotes the convolutional neural networks (CNNs) for corresponding inputs. Three CNNs all have f filters with latent dimensions of d , $2d$, and $3d$, respectively.

Then, to model the semantic information within the utterance, we propose a point-wise interaction between the current utterance $\mathbf{u}_t$ and each candidate slot-value pair $\mathbf{c}_\omega$ ($\omega \in \Omega$) via the following:

$$\mathbf{r}_{t,\omega}^{sem} = \mathbf{u}_t \otimes \mathbf{c}_\omega, \quad (3)$$

where $\mathbf{r}_{t,\omega}^{sem}$ is the semantic representation w.r.t. ω , and $\otimes$ is the element-wise product. Essentially, it learns the point-wise semantic similarity between $\mathbf{u}_t$ and $\mathbf{c}_\omega$, so $\mathbf{r}_{t,\omega}^{sem}$ is a vectorized representation on how well each slot-value pair aligns with the current utterance. Meanwhile, due to the limited amount of information in short sentences, the dialogue state cannot be fully expressed by a single user utterance. For example, when asked by the recommender whether the user likes Chinese food, the user may respond with only yes/no but does not imply other values (e.g., *American* or *French*) for slot *food type*. In such cases, no explicit slot-value pairs can be inferred from the user utterance, and the accuracy for dialogue state tracking will be largely affected. So, for each slot-value pair ω , we maintain the contextual information from previous conversation rounds with a representation $\mathbf{r}_{t,\omega}^{con}$:

$$\mathbf{r}_{t,\omega}^{con} = (\mathbf{d}_{t-1} \mathbf{c}_\omega) \mathbf{u}_t, \quad (4)$$

where $\mathbf{d}_{t-1}$ is the system response (see Section 3.2) generated at the previous conversation round $t - 1$, and the dot product of $\mathbf{c}_\omega$ and $\mathbf{d}_{t-1}$ generates a similarity score. A larger similarity score means that the candidate slot-value pair ω is more likely to keep the conversation coherent and vice versa. A larger similarity score means that the chosen candidate ω is more likely to keep the conversation coherent and vice versa. Then, the similarity score is used as a weighting factor to quantify the suitability of each candidate slot-value pair ω for utterance $\mathbf{u}_t$, represented by $\mathbf{r}_{t,\omega}^{con}$.

Based on the resulted semantic and contextual representations, namely, $\mathbf{r}_{t,\omega}^{sem}$ and $\mathbf{r}_{t,\omega}^{con}$, we further utilize a feed-forward network to fuse the information from both sides:

$$\mathbf{r}_{t,\omega} = \sigma(\mathbf{W}_{sem}\mathbf{r}_{t,\omega}^{sem} + \mathbf{W}_{con}\mathbf{r}_{t,\omega}^{con} + \beta_r), \quad (5)$$

where σ is the sigmoid activation function, $\mathbf{W}_{sem}$ and $\mathbf{W}_{con}$ are two weight matrices, β_r is the bias term, and $\mathbf{r}_{t,\omega}$ carries the information of both the current dialogue semantics and the contexts.

The dialogue state tracking approaches in one-way communication schemes [54, 55] only consider slot-value pairs from users' informing utterances (e.g., "I like Italian food") following a system-generated question (e.g., "What food type do you prefer"). So, it is hard to track users' abrupt additional requests/questions (e.g., "Is vegetarian food available in this restaurant"). Hence, to fully enable two-way communications, we feed each user utterance into a fine-grained *hierarchical decision process* to comprehensively track the dialogue state and respond to informing/requesting utterances accordingly.

With Equation (5), we can derive a specific representation $\mathbf{r}_{t,\omega}$ for each $\omega \in \Omega$. Afterwards, to generate the current dialogue state $\mathbf{b}_t$, we present a hierarchical decision process. In CRSAL, we not only compute the overall dialogue state $\mathbf{b}_t$ for the entire conversation, but also maintain a specific state $\mathbf{b}_{t,\omega}$ for each slot-value pair ω at conversation round t . Based on the previous $\mathbf{b}_{t-1,\omega}$, we update it using $\mathbf{r}_{t,\omega}$:

$$\mathbf{b}_{t,\omega} = softmax(\mathbf{W}_s\mathbf{b}_{t-1,\omega} + \mathbf{W}_r\mathbf{r}_{t,\omega}), \quad (6)$$

where $\mathbf{W}_s, \mathbf{W}_r \in \mathbb{R}^{2 \times d}$ are trainable matrices, and the $softmax(\cdot)$ is used to normalize the results into a binary probability distribution. Because $\mathbf{b}_{t,\omega} = [b_{t,\omega,1}, b_{t,\omega,2}]$ is a two-dimensional vector, a decision $\gamma_{t,\omega} \in \{0, 1\}$ on the *slot-value-level* is made for each ω :

$$\gamma_{t,\omega} = decision(\mathbf{b}_{t,\omega}) = \begin{cases} 0, & \text{if } b_{t,\omega,1} \leq b_{t,\omega,2} \\ 1, & \text{otherwise} \end{cases}. \quad (7)$$

Intuitively, this binary decision process judges whether the slot-value pair ω suits the users' demand (1 for positive and 0 for negative). Note that because $\omega = (s, v)$, $\gamma_{t,\omega} = \eta_{t,s,v}$. So, in the conversation round t , for each specific slot s , we further aggregate the binary results on different values v to generate *slot-level* decisions:

$$decision'(s) = [\gamma_{t,s,v_1}, \gamma_{t,s,v_2}, \dots, \gamma_{t,s,v_{n_s}}], \quad (8)$$

where n_s is the number of all candidate values for slot s . In short, $decision'(s)$ generates a one-hot encoding, where the most relevant values of slot s are marked with 1, and 0 otherwise. Finally, we generate the *conversation-level* representation of the current dialogue belief state $\mathbf{b}_t$, which is a one-hot vector constructed by concatenating all the slot-level decisions on s :

$$\mathbf{b}_t = decision'(s_1) \oplus decision'(s_2) \oplus \dots \oplus decision'(s_n), \quad (9)$$

where n is the total number of different slots. Similar to References [81, 89], our DST is pre-trained by minimizing the cumulative cross-entropy between the prediction $\hat{\mathbf{b}}_{t,\omega}$ and the ground truth label $\bar{\mathbf{b}}_{t,\omega}$,

$$\mathcal{L}_{DST} = - \sum_{\omega \in \Omega} (\hat{\mathbf{b}}_{t,\omega} \log(\bar{\mathbf{b}}_{t,\omega})), \quad (10)$$

where $\hat{\mathbf{b}}_{t,\omega}$ is the prediction of the belief state before the softmax.

3.1.2 Neural Intent Encoder. The user intent within each utterance is a user's purpose to deliver a specific type of information. For example, "I want an expensive restaurant" and "I like fine-dining" have the same intent of expressing preferences on high restaurant price range. Conventional methods mostly identify the intention through the predefined rules that are manually crafted and assign each intent an embedding vector [65, 77], which hinders both the flexibility and expressiveness of the modelled user intent. Thus, in CRSAL, inspired by the sequence-to-sequence learning framework [73], we introduce NIE, which encodes the word sequence derived from user utterance $\mathcal{U}_t = \{\mathbf{e}_m\}_{m=1}^M$ into a distributed user intent representation $\mathbf{z}_t$ that replaces the traditional hand-coded dialogue act representation [60]. Notably, utterances with similar intents are supposed to have intent representations that are close to each other and vice versa. To achieve this with NIE, we perform an initial delexicalization on the raw utterances, where the tokens representing the same slots/values are replaced by the same generic token. Meanwhile, to avoid the drawbacks of traditional predefined delexicalization approaches mentioned in Section 3.1.1, we adopt a similarity-based delexicalization on user utterances. Specifically, for each utterance token $\mathbf{u}_{t,m}$ within the synergic representation $\mathbf{u}_t = \mathbf{u}_{t,1} \oplus \dots \oplus \mathbf{u}_{t,M}$, we calculate its similarity to all $s \in \mathcal{S}$:

$$\begin{aligned} \varrho_{t,m} &= [\varrho_{t,m,s_1}, \varrho_{t,m,s_2}, \dots, \varrho_{t,m,s_n}] \\ &= [\sigma(\mathbf{u}_{t,m}^\top \mathbf{e}_{s_1}^{slo}), \sigma(\mathbf{u}_{t,m}^\top \mathbf{e}_{s_2}^{slo}), \dots, \sigma(\mathbf{u}_{t,m}^\top \mathbf{e}_{s_n}^{slo})], \end{aligned} \quad (11)$$

where $\mathbf{e}_s^{slo}$ is the embedding vector for slot s . Then, we select the best-matching slot s_* for token $\mathbf{u}_{t,m}$ with the following:

$$s_* = \begin{cases} s, & \text{if } \varrho_{t,m,s} \geq 0.9 \text{ and } \varrho_{t,m,s} = \max(\varrho_{t,m}) \\ NULL, & \text{otherwise} \end{cases}, \quad (12)$$

where we use 0.9 as the threshold to ensure the suitability of the identified slot s_* . Similarly, we select the best-matching value v_* from the corresponding value set $\mathcal{V}_{s_*}$ for slot s_* :

$$\begin{aligned} \xi_{t,m} &= [\xi_{t,m,v_1}, \xi_{t,m,v_2}, \dots, \xi_{t,m,v_{n_{s_*}}}] \\ &= [\sigma(\mathbf{u}_{t,m}^\top \mathbf{e}_{v_1}^{val}), \sigma(\mathbf{u}_{t,m}^\top \mathbf{e}_{v_2}^{val}), \dots, \sigma(\mathbf{u}_{t,m}^\top \mathbf{e}_{v_{n_{s_*}}}^{val})], \end{aligned} \quad (13)$$

and:

$$v_* = \begin{cases} v, & \text{if } \xi_{t,m,v} \geq 0.9 \text{ and } \xi_{t,m,v} = \max(\xi_{t,m}) \\ NULL, & \text{otherwise} \end{cases}, \quad (14)$$

where $\mathbf{e}_v^{val}$ is the embedding vector for value v .

We repeat Equation (11)–Equation (14) for all M tokens in $\mathbf{u}_t$. Then, if a slot/value s_*/v_* is identified in the m th utterance token, the corresponding representation $\mathbf{u}_{t,m}$ is replaced by the slot/value embedding $\mathbf{d}_{s_*}/\mathbf{d}_{v_*}$. Last, with the delexicalized utterance representation $\mathcal{U}'_t$, we feed the sequence into a bidirectional long short-term memory (biLSTM) network and concatenate the last forward and backward neural hidden state as the final representation of user intention:

$$\begin{aligned} \vec{\mathbf{z}}_t, \overleftarrow{\mathbf{z}}_t &= \text{biLSTM}(\mathcal{U}'_t), \\ \mathbf{z}_t &= \vec{\mathbf{z}}_t \oplus \overleftarrow{\mathbf{z}}_t. \end{aligned} \quad (15)$$

Essentially, the BiLSTM network effectively replaces the conventional hand-crafted intent representations by encoding all the latent semantic information of all tokens in the utterance into a dense representation. In Section 3.2, we will further introduce how $\mathbf{z}_t$ is used for conversational response generation.

3.1.3 Neural Recommender Network. In our conversational recommender system, to suggest suitable items to meet different user demands, we introduce our NRN that generates a ranked item list for each user according to the current dialogue state $\mathbf{b}_t \in \{0, 1\}^{|\Omega| \times 1}$. During the human-machine conversations, the user interacts with the conversational recommender turn-by-turn to express her/his demand on the ideal items. Assuming a recommendation response is decided by the system (see Section 3.2) in conversation round t , the NRN retrieves from the item database to rank all items for each user. The core idea is to utilize NRN to compute ranking scores for all items based on the extracted user preferences from dialogues and then form a ranking list based on the scores. For each item $i \in \mathcal{I}$, we use $\mathbf{f}_i \in \{0, 1\}^{|\mathcal{I}| \times 1}$ to denote its one-hot representation. As the dialogue state vector $\mathbf{b}_t$ encodes all slot-value pairs mentioned by the user, it carries substantial information about user preferences, so we concatenate it with the item one-hot vector $\mathbf{f}_i$ as the input of NRN:

$$\mathbf{x}_{t,i} = \mathbf{b}_t \oplus \mathbf{f}_i, \quad (16)$$

where $\mathbf{x}_{t,i} \in \{0, 1\}^{(|\Omega|+|\mathcal{I}|) \times 1}$ is the input feature vector. Now, the ranking score for item i can be generated by learning a regression function $f_{NRN}(\cdot)$ that maps $\mathbf{x}_{t,i}$ to the desired output. Since the input feature $\mathbf{x}_{t,i}$ in our conversational recommender system is high-dimensional but very sparse, a widely adopted solution is factorization machines (FMs) [61] that can learn cross-feature interactions. However, the relationships between the features and the predicted target in real application scenario is rather complex, and the FM models are much constrained by their limited expressiveness [23]. Hence, we design $f_{NRN}(\cdot)$ based on Reference [23], which is a non-linear extension of the plain FM, namely, the neural factorization machine with proven effectiveness on modelling sparse features and non-linear feature interactions:

$$f_{NRN}(\mathbf{x}_{t,i}) = w_0 + \sum_{j=1}^{|\Omega|+|\mathcal{I}|} w_j x_{t,i,j} + g(\mathbf{x}_{t,i}), \quad (17)$$

where the first two terms are the linear regression parts modelling global bias of data and weight of the features [61], while $x_{t,i,j}$ and w_j , respectively, denote the j th element of vector $\mathbf{x}_{t,i}$ and the weight assigned to it. Function $g(\cdot)$ in the third term is a deep neural network modelling the non-linear feature interactions.

The design of $g(\cdot)$ is shown in Figure 5. The embedding layer projects each input feature element $x_{t,i,j}$ into a dense vector representation $\mathbf{v}_j$, resulting in a set of embedding features $\mathcal{X} = \{x_{t,i,1}\mathbf{v}_1, \dots, x_{t,i,|\mathcal{I}|}\mathbf{v}_{|\mathcal{I}|}\}$. On top of the embedding layer, we utilize a bi-interaction layer [23] as a pooling function to fuse the embedded features into a unified one:

$$f_{BI}(\mathcal{X}) = \sum_{j=1}^{|\mathcal{I}|} \sum_{j'=j+1}^{|\mathcal{I}|} x_{t,i,j} \mathbf{v}_j \otimes x_{t,i,j'} \mathbf{v}_{j'}, \quad (18)$$

where the bi-interaction operation is denoted by $f_{BI}(\cdot)$. Intuitively, it encodes the second-order interactions between features into a compact vector representation. Then, with a single-layer feed-forward network and a linear projection, we can complete the output of $g(\cdot)$:

$$g(\mathbf{x}_{t,i}) = \mathbf{w}_g^\top \sigma(\mathbf{W}_g f_{BI}(\Omega_x) + \beta_g), \quad (19)$$

where σ is the sigmoid activation function, $\mathbf{W}_g$ and $\mathbf{w}_g$ are, respectively, network and projection weights, while β_1 is the network bias term. So far, we can rewrite Equation (17) as:

$$f_{NRN}(\mathbf{x}_{t,i}) = w_0 + \sum_{j=1}^{|\Omega|+|\mathcal{I}|} w_j x_{t,i,j} + \mathbf{w}_g^\top \sigma(\mathbf{W}_g f_{BI}(\mathcal{X}) + \beta_g). \quad (20)$$

Fig. 5. The non-linear part of the Neural Factorization Machine.

With the fully trained NRN, it ranks all candidate items according to the score $f_{NRN}(\mathbf{x}_{t,i})$ in conversation round t and return the ranking list as the recommendation results. Also, we pre-train the NRN and freeze its parameters when training the entire CRSAL model. Considering the existing datasets for conversational recommender system and task-oriented dialogue system only contain positive instances (i.e., all instance have target value 1), we need to sample some negative instances to pair with one positive instance to ensure the generalization of the predictive model. In our case, we randomly alter two dialogue belief state values, which belong to different slot names as the negative instances. Three negative instances are coupled with one positive instance to construct the training set. The model training objective is to minimize the squared loss between the prediction and the target,

$$\mathcal{L}_{NRN} = - \sum_{i \in I} (f_{NRN}(\mathbf{x}_{t,i}) - f_{ng}(\mathbf{x}_{t,i}))^2, \quad (21)$$

where $f_{ng}(\mathbf{x}_{t,i})$ denotes the target of input $\mathbf{x}_{t,i}$, and we update the model parameters with stochastic gradient descent method. Without loss of generality, to the best of our knowledge, this is the first time that the dialogue state and item information are jointly modelled in a non-linear way for conversational recommendation tasks.

3.2 Conversational Response Generation

After the information extraction stage, the resulted information is passed into the conversational response generation module to generate dialogue reactions. In general, the conversational response generation module first decides which system action to take (e.g., greet the user, ask for more information, generate recommendations) and then composes several sentences to complete the chosen action (e.g., asking “what is your budget” to require additional information). Correspondingly, in this module, a policy agent is built to decide the system action in each conversation round t conditioned on the extracted information (i.e., dialogue state $\mathbf{b}_t$ and user intention $\mathbf{z}_t$). After a system action is decided, we further develop a language generator to

Fig. 6. The overview of the conversational response generation module.

generate the conversational response with the guidance of the action. In traditional policy agents, the system actions are manually designed by experts with extensive domain knowledge, which limits the expressiveness of action representations [98]. Besides, existing task-oriented dialogue systems suffer from insufficiency of training corpus, which leads to the problems of overfitting and generalization [81] when being applied to more generic recommendation scenarios.

To solve the aforementioned problems, inspired by the prior research on neural variational inference [8, 34, 64], we jointly design a neural policy agent (NPA) and a natural language generator (NLG) in the conversational response generation module for learning discourse-level latent representations of actions and guiding the conversation in more sophisticated domains. An overview of our proposed conversational response generation module is shown in Figure 6. We assume that utterance representing the same dialogue action will share many common keywords. Furthermore, a new type of latent dialogue action representation is proposed with the assumption that the utterances representing the same dialogue action will share many common words. In short, we adopt a clustering-based mechanism to infer the dialogue context and generate the latent system action representations for each utterance.

3.2.1 Neural Policy Agent. First, in NPA, to explicitly quantify the amount of information received from a user's utterances, we define an information richness indicator $\mathbf{q}_t \in \{0, 1\}^{d_q \times 1}$ for conversation round t . Suppose the total amount of different slots mentioned till round t is η_t , then each element $q_{t,j} \in \mathbf{q}_t$ is defined as:

$$q_{t,j} = \begin{cases} 1, & \text{if } j = \eta_t \text{ or } j = d_q < \eta_t, \\ 0, & \text{otherwise} \end{cases}, \quad (22)$$

where d_q is the total dimension of all slots. Given the dialogue belief state $\mathbf{b}_t$ and user intention encoding $\mathbf{z}_t$, we first construct the internal dialogue state vector $\mathbf{s}_t$ with:

$$\mathbf{s}_t = \mathbf{q}_t \oplus \mathbf{b}_t \oplus \mathbf{z}_t, \quad (23)$$

and then formulate the probability distribution of system actions a_t using a multi-layer perceptron (MLP) unit:

$$\pi_{\Theta}(a_t | \mathbf{s}_t) = \text{softmax}(\mathbf{W}_2 \cdot \tanh(\mathbf{W}_1 \mathbf{s}_t + \beta_1) + \beta_2), \quad (24)$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, β_1 , and β_2 are parameters to learn, and Θ denotes all trainable parameters in the NPA. Based on this conditional distribution, we can sample a response action a_t (i.e., the dialogue

action in the conventional task-oriented dialogue system) for the current dialogue state:

$$\hat{a}_t \sim \pi_\Theta(a_t | s_t). \quad (25)$$

3.2.2 Natural Language Generator. The sampled dialogue action $\hat{a}_t$ and the dialogue internal state s_t are then utilized by our NLG for obtaining a dialogue control vector $\mathbf{o}_t$:

$$\mathbf{o}_t = \mathbf{W}_5 \mathbf{a}_t \oplus (\sigma(\mathbf{W}_3 \mathbf{a}_t + \beta_3) \cdot \mathbf{W}_4 \mathbf{s}_t), \quad (26)$$

where $\mathbf{a}_t$ is the one-hot encoding for the sampled action $\hat{a}_t$, $\mathbf{W}_3$, $\mathbf{W}_4$, $\mathbf{W}_5$, and β_3 are trainable parameters. Intuitively, the degree of information flow from the internal dialogue state s_t is controlled by a sigmoid gate whose input signal is the sampled system action $\mathbf{a}_t$. Next, the control vector $\mathbf{o}_t$ will govern the generation of conversational response $\mathcal{D}_t = \{\tau_{t,1}, \tau_{t,2}, \dots, \tau_{t,K}\}$ using a conditional LSTM [81] language model:

$$p_\Theta(\mathcal{D}_t | a_t, s_t) = \prod_{k \in \{1, \dots, K\}} p(\tau_{t,k+1} | \tau_{t,k}, \mathbf{h}_{t,k-1}, \mathbf{o}_t), \quad (27)$$

with $\tau_{t,k}$ as the k th output token (i.e., a word, a delexicalized slot name, or value), and $\mathbf{h}_{t,k-1}$ the LSTM decoder's last hidden state, and the $\mathcal{D}_t$ is the conversational response in the t th conversation turn consisting of a sequence of tokens. We can obtain the parameterized model using internal state s_t and the response $\mathcal{D}_t$ as the following:

$$p_\Theta(\mathcal{D}_t | s_t) = \sum_{\forall a_t} p_\Theta(\mathcal{D}_t | a_t, s_t) \pi_\Theta(a_t | s_t). \quad (28)$$

3.2.3 Parameter Inference. To conduct parameter inference, we need to approximate the posterior distribution of the latent action $p_\Theta(a_t | s_t, \mathcal{D}_t)$. Hence, we learn an inference network $y_\Phi(a_t | s_t, \mathcal{D}_t)$ so that we can optimize the lower bound of the joint probability in the neural variational inference framework [50]. The lower bound of the variational inference is the following log-likelihood:

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{y_\Phi(a_t | s_t, \mathcal{D}_t)} [\log p_\Theta(\mathcal{D}_t | a_t, s_t)] - d_{KL}(y_\Phi(a_t | s_t, \mathcal{D}_t) \parallel \pi_\Theta(a_t | s_t)) \\ &\leq \log \sum_{a_t} p_\Theta(\mathcal{D}_t | a_t, s_t) \pi_\Theta(a_t | s_t) \\ &= \log p_\Theta(\mathcal{D}_t | s_t), \end{aligned} \quad (29)$$

where we deploy a trade-off factor $\lambda > 0$ on the Kullback-Leibler (KL) divergence $d_{KL}(\cdot)$ when computing the loss:

$$\mathcal{L} = \mathbb{E}_{y_\Phi(a_t | s_t, \mathcal{D}_t)} [\log p_\Theta(\mathcal{D}_t | a_t, s_t)] - \lambda d_{KL}(y_\Phi(a_t | s_t, \mathcal{D}_t) \parallel \pi_\Theta(a_t | s_t)). \quad (30)$$

Hereby, the inference network $y_\Phi(a_t | s_t, \mathcal{D}_t)$ can be formulated as:

$$\begin{aligned} y_\Phi(a_t | s_t, \mathcal{D}_t) &= \text{softmax}(\mathbf{W}_e \mathbf{e}_t^{all}), \\ \mathbf{e}_t^{all} &= \sigma(\mathbf{W}_b \mathbf{b}_t + \mathbf{W}_q \mathbf{q}_t + \mathbf{W}_z \mathbf{z}_t + \mathbf{W}_d \mathbf{d}_t + \beta_y), \end{aligned} \quad (31)$$

where $\mathbf{W}_e$, $\mathbf{W}_b$, $\mathbf{W}_q$, $\mathbf{W}_z$, $\mathbf{W}_d$, and β_y are corresponding trainable parameters. We use $\mathbf{e}_t$ here to represent the combination of dialogue belief state $\mathbf{b}_t$, information richness indicator $\mathbf{q}_t$, user intention encoding $\mathbf{z}_t$, and the encoding of system-generated conversation $\mathbf{d}_t$. In the inference network $y_\Phi(a_t | s_t, \mathcal{D}_t)$, it is worth noting that the representation of the system-generated dialogue

$\mathbf{d}_t$ is calculated by encoding the token sequence $\mathcal{D}_t = \{\tau_{t,1}, \tau_{t,2}, \dots, \tau_{t,K}\}$ into a single vector using biLSTM with an embedding layer:

$$\begin{aligned} \vec{\mathbf{d}}_t, \overleftarrow{\mathbf{d}}_t &= \text{biLSTM}(\mathcal{D}_t), \\ \mathbf{d}_t &= \vec{\mathbf{d}}_t \oplus \overleftarrow{\mathbf{d}}_t. \end{aligned} \quad (32)$$

In our policy agent, although the $\pi_\theta(a_t|s_t)$ and $y_\phi(a_t|s_t, \mathcal{D}_t)$ are the parameterized distribution w.r.t. the dialogue action a_t , the approximation $y_\phi(a_t|s_t, \mathcal{D}_t)$ only functions during inference by producing samples to compute the stochastic gradients, while $\phi_\theta(\mathcal{D}_t|s_t)$ is responsible for generating the action samples to guide the generation of conversational responses. Based on the sampled system actions from the inference network $a_t \sim y_\phi(a_t|s_t)$, we can optimize the parameters Θ and Φ against the variational lower bound. In what follows, we describe our different optimization strategies specifically designed for the NPA and NLG based on Equation (28).

For parameter set Θ , we divide it into two parts, i.e., $\Theta = \{\Theta_{NPA}, \Theta_{NLG}\}$. In terms of Θ_{NPA} in the policy agent, we use the back-propagation to update the parameters:

$$\nabla_{\Theta_{NPA}} \mathcal{L} = \nabla_{\Theta_{NPA}} \mathbb{E}_{y_\phi(a_t|s_t, \mathcal{D}_t)} [\log p_{\Theta_{NPA}}(\mathcal{D}_t|a_t, s_t)]. \quad (33)$$

As for the parameters Θ_{NLG} in the NLG, they are updated by minimizing the KL divergence:

$$\nabla_{\Theta_{NLG}} \mathcal{L} = -\nabla_{\Theta_{NLG}} \lambda d_{KL}(y_\phi(a_t|s_t, \mathcal{D}_t) || \pi_{\Theta_{NLG}}(a_t|s_t)). \quad (34)$$

Meanwhile, we update the parameters Φ within the inference network $y_\phi(a_t|s_t, \mathcal{D}_t)$ by first defining the following learning signal $r(\mathcal{D}_t, a_t, s_t)$:

$$r(\mathcal{D}_t, a_t, s_t) = \log p_{\Theta_{NPA}}(\mathcal{D}_t|a_t, s_t) - \lambda(\log y_\phi(a_t|s_t, k_t)) - \log(\pi_{\Theta_{NLG}}(a_t|s_t)) + b_r, \quad (35)$$

with the scalar b_r to adjust the estimation variance. Then, we can use the learning signal to update Φ via:

$$\nabla_\Phi \mathcal{L} = \nabla_\Phi \mathbb{E}_{y_\phi(a_t|s_t, \mathcal{D}_t)} [r(\mathcal{D}_t, a_t, s_t) \log y_\phi(a_t|s_t, \mathcal{D}_t)]. \quad (36)$$

An unsupervised learning scheme will be subject to the high variance of the samples from the inference network and the overly strong influence of the NLG $p_\Theta(\mathcal{D}_t|a_t, s_t)$. So, to stabilize the training process, we adopt a weighted semi-supervised learning method to determine the system actions in the latent space. Inspired by Reference [71], we cluster the latent actions of the utterance based on the keywords in the delexicalized utterance, thus automatically generating the label $\bar{a}_t$ for each training example. As such, we can obtain a small portion of labelled training instances $(\mathcal{D}_t, s_t, \bar{a}_t) \in \bar{\mathcal{T}}$. In $\bar{\mathcal{T}}$, we treat $\bar{a}_t$ as the observed variable and maximize the joint log-likelihood to train the model:

$$\mathcal{L}_1 = \sum_{(\mathcal{D}_t, s_t, \bar{a}_t) \in \bar{\mathcal{T}}} \log[p_{\Theta_{NPA}}(\mathcal{D}_t|\bar{a}_t, s_t) \pi_{\Theta_{NLG}}(\bar{a}_t|s_t) y_\phi(\bar{a}_t|s_t, \mathcal{D}_t)]. \quad (37)$$

Furthermore, when the model is trained on the unlabelled instances $(\mathcal{D}_t, s_t) \in \mathcal{T}$, we optimize it against the variational lower bound, which is similar to Equation (28):

$$\mathcal{L}_2 = \sum_{(\mathcal{D}_t, s_t) \in \mathcal{T}} \mathbb{E}_{y_\phi(a_t|s_t, \mathcal{D}_t)} [\log p_{\Theta_{NPA}}(\mathcal{D}_t|a_t, s_t)] - \lambda d_{KL}(y_\phi(a_t|s_t, \mathcal{D}_t) || \pi_{\Theta_{NLG}}(a_t|s_t)). \quad (38)$$

At last, the final loss $\mathcal{L}^*$ is formulated as the following:

$$\mathcal{L}^* = \mathcal{L}_1 + \eta \mathcal{L}_2, \quad (39)$$

where we use a parameter η to balance the influence of the supervised and unsupervised losses.

Fig. 7. The overview of the proposed adversarial Advantage Actor-Critic reinforcement learning.

3.3 Adversarial Reinforcement Learning

In Section 3.2, we have learned a neural policy agent $\pi_\theta(a_t|s_t)$ that infers a system action a_t in each conversation round to guide the conversation generation. However, the limited training corpus makes it hard to derive a satisfying policy agent for completing conversational recommendation. Besides, the biases within end-users' feedback and its insufficiency make it hard to be used as the reward signal to fine-tune our NPA. Therefore, we introduce an adversarial reinforcement learning approach illustrated in Figure 7 as a solution to train the NPA.

The decision procedure of a dialogue action in a conversation can be regarded as a partially observable Markov decision process (POMDP). The policy agent should select the best system action in each conversation round so that the long-term objective can be optimized. The Actor-Critic-based (AC-based) methods have shown promising performance compared with other reinforcement learning (RL) methods for dialogue policy decision-making [57, 70]. AC-based methods assume that a dialogue system behaving like human experts could lay positive influence on long-term performance and adopt an *actor* to generate an action to imitate human behaviors and then leverage a *critic* to judge whether the generated response is helpful to the final objective or not. The output of the critic serves as the reward signal to regulate the actor to generate high-quality system actions. Recently, an advantage Actor-Critic (A2C) method is proposed, where an advantage function is designed to decrease the cumulative variance in the reward signal and stabilize the RL process [22, 51]. Inspired by the latest generative adversarial networks (GANs) [20], we additionally design a discriminator in our system to differentiate the system-generated responses from the ones made by human experts, whose output could be used as an intrinsic reward in the A2C framework. The intuition behind this method is that, if the system-generated actions are close to the high-quality actions made by humans, then those generated actions can further help generate high-quality dialogues, so a high reward will be given. In this way, the embedded adversarial learning in the advantage Actor-Critic framework can encourage our NPA to generate more realistic and human-like actions during the conversation.

To this end, we propose a novel solution to perform adversarial reinforcement learning on the neural policy agent, namely, the adversarial advantage Actor-Critic (A3C) framework. In A3C, the neural policy agent (NPA) $\pi_\theta(a_t|s_t)$ serves as the actor to generate system action a_t conditioned on current dialogue state s_t , the critic computes an advantage function $A_{\pi_\theta(a_t|s_t)}(s_t, a_t)$ quantifying the benefit of the current dialogue action a_t , and the introduced discriminator $\tau_\phi(s_t, a_t)$ is trained to differentiate the system-generated actions from the ground truth responses made by humans. The training objective is to find a proper policy agent that maximizes the expected dialogue reward $R = \sum_{t=0}^{T-1} r_t$ over all candidate dialogue trajectories, where T is the length of the conversation and r_t is the RL reward signal at time stamp t . Based on the NPA $\pi_\theta(a_t|s_t)$ and reward $Q_{\pi_\theta(a_t|s_t)}(s_t, a_t)$,

we can obtain the gradient of the parameters in NPA with:

$$\nabla_{\Theta} J(\Theta) = \nabla_{\Theta} \mathbb{E}[\pi_{\Theta}(a_t | s_t) Q_{\pi_{\Theta}(a_t | s_t)}(s_t, a_t)]. \quad (40)$$

Because the reward signal relies on the action a sampled by NPA $\pi_{\Theta}(a_t | s_t)$, it may cause a large variance on the reward signal, which can weaken the training robustness. A baseline function $\psi(s)$ is studied in Reference [51] to reduce the variance while keeping the estimated gradient unchanged. Thus, we design a baseline function $\psi(s) = FFN(s)$ to represent the current state value with a single layer feedforward neural network. By constructing an advantage function with the reward signal and baseline function, we can quantify the impact of the current action to the task completion, which also reflects the quality of the system-generated action. As such, Equation (40) is rewritten as:

$$\begin{aligned} \nabla_{\Theta} J(\Theta) &= \nabla_{\Theta} \mathbb{E}[\log \pi_{\Theta}(a_t | s_t) A_{\pi_{\Theta}(a_t | s_t)}(s_t, a_t)], \\ A_{\pi_{\Theta}(a_t | s_t)}(s_t, a_t) &= Q_{\pi_{\Theta}(a_t | s_t)}(s_t, a_t) - \psi(s_t). \end{aligned} \quad (41)$$

However, we have to train two groups of parameters in this setting. To simplify the mathematical form of the advantage function and increase model stability, we leverage temporal difference (TD) error as an unbiased estimation of the advantage function:

$$\delta_{\pi_{\Theta}(a_t | s_t)} = r_t(s_t, a_t) + \psi(s_{t+1}) - \psi(s_t), \quad (42)$$

where s_{t+1} is the updated dialogue state in the next conversation round. Such an advantage function gives the same weight for each turn in the conversation, which can be problematic, as the actor tends to prolong the conversation instead of finishing the task. Here, we apply a damping factor $\gamma \in [0, 1)$ on $\psi(s_{t+1})$ to encourage the model to finish the task in fewer conversation rounds. The modified TD error can be written as the following:

$$\delta_{\pi_{\Theta}(a_t | s_t)} = r_t(s_t, a_t) + \gamma \psi(s_{t+1}) - \psi(s_t). \quad (43)$$

So far, the gradient function can be updated as:

$$\nabla_{\Theta} J(\Theta) = \nabla_{\Theta} \mathbb{E}[\log \pi_{\Theta}(a_t | s_t) \delta_{\pi_{\Theta}(a_t | s_t)}]. \quad (44)$$

In our case, as the actor is responsible for generating human-like system actions, the role of the actor can be regarded as the generator in our adversarial learning framework. Meanwhile, the discriminator is designed to distinguish the generated actions with the ground truth actions made by humans:

$$\begin{aligned} \rho_t &= LSTM(s_t, a_t), \\ \tau_{\phi}(s_t, a_t) &= \sigma(\mathbf{W}_r \rho_t + \beta_r), \end{aligned} \quad (45)$$

where σ is the logistic *Sigmoid* function, $\mathbf{W}_r$ and β_r are the weight and bias in the discriminator's output layer. The output of the discriminator serves as an intrinsic reward to force the generator to produce human-like response actions. We require the discriminator to be able to accurately distinguish the generated actions from the NPA from human-made ones, so that the NPA can be fully trained and eventually react like a human. We can formulate this training objective as a minimax game to find the parameters at saddle point:

$$J_{\pi^*, \tau^*} = \min_{\Theta} \max_{\phi} \mathbb{E}_{(s, a) \sim \pi_{\text{true}}(a | s)} [\log \tau_{\phi}(s_t, a_t)] + \mathbb{E}_{(s, a) \sim \pi_{\Theta}(a | s)} [\log(1 - \tau_{\phi}(s_t, a_t))]. \quad (46)$$

The objective of our discriminator is to maximize the likelihood of successfully differentiating the system-generated actions and human actions. We treat the generated actions as negative samples while the human-made ones are treated as positive samples, and the discriminator is trained

to distinguish them. The optimal parameter for the discriminator is:

$$\phi^* = \max_{\phi} \mathbb{E}_{(s,a) \sim \pi_{\text{true}}(a|s)} [\log \tau_{\phi}(s_t, a_t)] + \mathbb{E}_{(s,a) \sim \pi_{\Theta^*}(a_t|s_t)} [\log(1 - \tau_{\phi}(s_t, a_t))], \quad (47)$$

The output of $\tau_{\phi}(s, a)$ is regulated with the range of $[0, 1]$, which can be viewed as a probability indicating how similar the generated action is to the human-made action. This metric serves as an important intrinsic reward to construct the reinforcement learning reward $r_t(s_t, a_t)$ in Equation (43):

$$r_t(s_t, a_t) = \tau_{\phi}(s_t, a_t) + r^{ben}. \quad (48)$$

For balancing the dialogue quality and task completion, we add another intermediate metric r^{ben} , which indicates whether the generated dialogue action is beneficial for finishing the task, into the reward function. Here, we determine whether a dialogue action is beneficial by observing the impact of the system-generated action to task completion:

$$r^{ben} = \begin{cases} 1, & \text{the } a_t \text{ has a positive influence} \\ -1, & \text{the } a_t \text{ has a negative influence.} \\ 0, & \text{otherwise} \end{cases} \quad (49)$$

Algorithm 1 presents the training process based on our proposed A3C framework. Though there is a recent enhancement on A2C methods using adversarial learning [51], it directly adopts RL to train the whole dialogue system from scratch, and the model parameters require a two-phase update procedure, i.e., the pure A2C process and the A2C process with adversarial learning. Squeezing these two RL steps into one training batch inevitably increases the difficulty in finding a set of universally optimal parameters, and it is resource-consuming. In contrast, we first pre-train the policy agent in CRSAL and then fine-tune it in our proposed A3C scheme, hence the discriminator can directly measure the similarity between the generated response and human-made ones for calculating the RL reward, making it faster and easier to obtain the optimal model parameters.

4 EXPERIMENTS

4.1 Datasets

We implement CRSAL on three public datasets, namely, DSTC2, CamRest676, and MultiWOZ 2.1, to verify its effectiveness and performance on conversational recommendation. The first two datasets focus on the task of helping users find restaurants that meet their requirements in Cambridge, UK, and they share the same ontology and restaurant database. The third dataset is currently one of the most challenging multi-domain human-human dialogue datasets, having seven different daily conversation scenarios. We also select the dialogues in the restaurant domain to perform conversational recommendation. These three datasets are generated by human participants and further preprocessed by human experts, making them fully align with real human-human conversational interactions. The details of the three datasets are introduced below:

DSTC2 [24]: DSTC2 is a dataset collected via automatic speech recognition in the restaurant searching domain. Due to the complex acoustic conditions, the raw dataset inevitably contains many linguistic errors [55]. In our experiment, we adopt a manually cleaned version of DSTC2.¹ We follow the same dividing strategy for training, validation, and test as Reference [24]. The training set includes 1,612 conversations, while there are 506 and 1,117 conversations, respectively, in the validation and test set. In each conversation session, the user can use three kinds of informing slots (i.e., *price range*, *food type*, *area*)

¹<https://www.dropbox.com/s/eqv2vgc8labxk0r/dstc2.rar?dl=0>.

ALGORITHM 1: Adversarial Advantage Actor-Critic Reinforcement Learning

Input: Ground truth training actions $\mathcal{A}$, neural policy agent $\pi_{\Theta}(a_t|s_t)$, discriminator $\tau_{\phi}(s_t, a_t)$, baseline function $\psi(s_t)$, maximum training iteration N ;

Result: Fine-tuned neural policy agent $\pi_{\Theta}(a|s)$;

- 1 Pre-train the neural policy agent $\pi_{\Theta}(a|s)$ based on the given training corpus $\mathcal{A}$;
- 2 Use pre-trained neural policy agent $\pi_{\Theta}(a|s)$ to generate a set of simulated system actions $\mathcal{A}^{NPA}$;
- 3 Draw positive samples $\mathcal{A}^+$ and negative samples $\mathcal{A}^-$, respectively, from $\mathcal{A}$ and $\mathcal{A}^{NPA}$;
- 4 Pre-train the discriminator $\tau_{\phi}(s_t, a_t)$ with $\mathcal{A}^+$ and $\mathcal{A}^-$;
- 5 $N \leftarrow 0$;
- 6 **while** $i \leq N$ **do**
- 7 $buffer \leftarrow \emptyset$;
- 8 **while** t is not the final conversation turn **do**
- 9 Generate system action a_t with policy agent $\pi_{\Theta}(a|s)$ and current dialogue state s_t ;
- 10 Calculate the reward r_t with Equation (48);
- 11 Obtain a new state s_{t+1} for conversation round $t + 1$ with Equation (23);
- 12 $buffer \leftarrow (s_t, a_t, r_t, s_{t+1})$;
- 13 **end**
- 14 Compute the TD error with Equation (43);
- 15 Take a gradient step to optimize neural policy agent $\pi_{\Theta}(a_t|s_t)$ with Equation (44);
- 16 Take a gradient step to optimize discriminator $\tau_{\phi}(s_t, a_t)$ with Equation (47);
- 17 **end**

to clarify her/his demand, which are identified by our conversational recommender system to generate the recommendation. Once the user has selected a restaurant, she/he can further use seven inquiring slots (i.e., *name*, *phone number*, *address*, *postcode* plus three informing slots) to ask for further information about the restaurant.

CamRest676 [82]: The CamRest676 dataset is collected based on a modified Wizard-of-Oz paradigm [33], and it shares the same restaurant information with DSTC2. Human participants were recruited on Amazon Mechanical Turk and asked to compose the conversations from the perspective of either a user or a wizard. During the data collection process, a user is asked to follow a pre-specified user goal to interact with the wizard, so the collected dialogue is much closer to reality. There are 676 dialogues in the dataset. We use the same dataset preprocessing method used in Reference [82] and split the dataset for training, validation, and test with the ratio of 3:1:1.

MultiWOZ 2.1 [18]: The MultiWOZ 2.1 is an updated version of MultiWOZ 2.0 dataset, which is one of the largest publicly available multi-domain human-human dialogue datasets. This dataset is also collected via the Wizard-of-Oz paradigm, which guarantees that the training corpus is close to the real-world scenarios. In MultiWOZ 2.1, there are seven domains, 13 intents, 25 slot categories, and 10,483 dialogue sessions. We select the restaurant reservation domain, which includes 4,688 dialogues and 402 different slot values to validate the conversational recommendation effectiveness. It is a challenging benchmark dataset for evaluating the task-oriented dialogue systems due to its inherent topic diversity and rich linguistic variations. Even though it still provides three informing slots as in the first two datasets, it contains substantially more slot values for users to describe their fine-grained demands. Also, this dataset includes more dialogue actions (e.g., informing the number of attendance, making a reservation, asking for reference number, cancelling the booking).

Note that these three datasets contain anonymized users (i.e., user IDs are unknown) and positive instances only (i.e., an item is successfully recommended at the end of each conversation), so we reconstruct the data by sampling negative instances to train all the conversational recommender systems. Because the conversation dataset is anonymous, users' historical interactions with items and their side information are unavailable. So, the recommendation problem in our system is a classic cold start problem [91], where the recommender systems generate the recommendation results based on the collected conversational interactions without any knowledge on users' previous behaviors or personal profiles. To achieve a successful conversational recommendation, the conversational recommender systems should be able to generate suitable dialogue responses during the conversation and accurately find users' desired items simultaneously.

4.2 Parameter Settings

For the neural components in our proposed CRSAL, the latent dimension of all hidden layers and embedding layers is set to 50, while all the weights are randomly initialized in $[-0.3, 0.3]$. The model is trained using ADAM optimizer [35] with a learning rate of 0.001 for DST, 0.003 for NRN and LVILM, and 0.008 for the A3C-based RL module. The clustering dimension H of the latent actions is set to 70, which means that we treat the first 70 response clusters as the supervised training set $\overline{\mathcal{T}}$, where an additional dictionary for functional words is applied to enhance the clustering accuracy. The pre-trained *Paragram-SL999* word vectors [52] are used for producing the utterance embedding [83]. Also, the trade-off parameter η in the semi-supervised learning is set to 0.1, and we use gradient clipping in $[-0.2, 0.2]$ to prevent gradient explosion and stabilize the training process.

4.3 Evaluation Metrics

The ultimate goal of the proposed CRSAL is to recommend suitable items to users through human-like conversations. We assess the effectiveness of the CRSAL from three perspectives: the quality of conversation generation, the accuracy of items retrieved, and the success rate of recommendation. Correspondingly, we leverage four evaluation metrics, namely, BLEU, ROUGE, entity matching rate (EMR), and success rate, which are described as follows:

- BLEU:** BLEU is a widely used metric for machine translation and task-oriented dialogue tasks. It evaluates the quality of generated conversational responses by calculating the co-occurrences of n -grams in the ground truth and the proposed responses [46, 56]. The BLEU score is the most direct way to indicate the similarity of the generated responses to the reference sentences.
- ROUGE:** ROUGE-L (ROUGE for short) is a standard metric for evaluating the performance of the language generator/translator. It measures how well the words in the generated response and the ground truth align with each other. ROUGE is calculated by measuring the longest common subsequence of the generated sentences and the ground truth sentences [2, 42].
- EMR:** EMR is a metric that indicates whether the items retrieved by the conversational recommender system at the end of the conversation match the ground truth (i.e., the items actually accepted by users) [24, 82]. The EMR score essentially reflects the accuracy of the retrieval.
- SR:** SR evaluates the ratio of successfully completed tasks. A conversational recommendation task is successful only if: (1) the recommended items match a user's demand; (2) the conversational recommender system can correctly react to a user's further queries about

the recommended item [82] and respond with the right information (e.g., the address of a restaurant).

In addition, we propose a synergic evaluation metric in our experiments for conversational recommendation, which evaluates the overall performance from both recommendation and conversation perspectives. We term this joint metric quality of conversational recommendation QoCR, and it is defined as:

$$QoCR = w_{SR} \cdot SR + w_{BLEU} \cdot BLEU, \quad (50)$$

where the BLEU and SR can be balanced with w_{SR} and w_{BLEU} for different research goals. As our ultimate target is to ensure a successful delivery of recommendation with the assistance of conversation, we set $w_{SR} = 1$ and $w_{BLEU} = 0.5$ in our experiments.

4.4 Baselines

To better demonstrate the performance of the proposed system, we select several state-of-the-art baselines for comparison. They are summarized as follows:

- CRS:** Conversational recommender system (CRS) [72] is a dialogue system for recommending items to the user. It involves a semi-structured user query to represent the dialogue history as well as the user intent information. This query is then used to make a personalized recommendation. The dialogue agent is guided by a policy gradient-based reinforcement learning policy network to generate suitable conversational responses. The main structure is based on the traditional task-oriented dialogue system [97] with a recommender system added to it.
- NDM:** The network-based dialogue model (NDM) [82] is an end-to-end task-oriented dialogue system. It is the first work to propose a unified dialogue structure that combines belief tracking, knowledge base operation, and response generation together. This model could track the dialogue belief state, inference with a knowledge base, and incorporate structured knowledge base query into system responses to complete the task.
- LIDM:** The latent intention dialogue model (LIDM) is a pioneer in using the variational inference network to derive the parameters of the policy agent and language generator. It adopts the evidence lower bound (ELBO) [48] to approximate the posterior distribution of the dialogue action, which could fully uncover the latent action space and generate accurate action prediction. The introduced latent action representations perform better compared with the explicit dialogue action derived from the sentence action classification [81] on response generation and task completion. It also gives a parameter optimization method with gradient-based reinforcement learning.
- MLEM:** The maximum likelihood estimation model (MLEM) is a variant of LIDM [89]. It adopts the log-likelihood method instead of ELBO to derive the distribution of the dialogue action. It spares more attention onto natural language response generation by sacrificing the accuracy on action prediction.
- TSCP:** This method introduces a two-stage copynet (TSCP) framework to build a task-oriented dialogue system, which can handle both task-completion and response generation in a single Seq2Seq model. TSCP first uses a belief span method to track the dialogue states, which will further serve as the input source for Seq2Seq modelling [37]. The copynet incorporates the belief span value into the final response generation to solve the out-of-vocabulary problem in the task-oriented dialogue system.

Table 1. Effectiveness Results on Three Datasets

Dataset	Method	BLEU	ROUGE	EMR	SR	QoCR
DSTC2	CRS	0.210	0.342	0.695	0.603	0.708
	NDM	0.208	0.327	0.702	0.582	0.686
	LIDM	0.224	0.382	0.709	0.621	0.733
	MLEM	0.235	0.414	0.707	0.608	0.726
	TSCP	0.204	0.301	0.697	0.605	0.707
	CRSAL	0.230	0.395	0.719	0.628	0.743
CamRest676	CRS	0.228	0.389	0.885	0.791	0.905
	NDM	0.224	0.371	0.904	0.779	0.891
	LIDM	0.248	0.475	0.916	0.838	0.962
	MLEM	0.304	0.503	0.894	0.819	0.971
	TSCP	0.225	0.374	0.917	0.818	0.931
	CRSAL	0.262	0.488	0.925	0.854	0.985
MultiWOZ 2.1	CRS	0.146	0.267	0.587	0.435	0.508
	NDM	0.181	0.328	0.721	0.513	0.604
	LIDM	0.207	0.352	0.764	0.572	0.676
	MLEM	0.195	0.337	0.733	0.536	0.634
	TSCP	0.165	0.291	0.656	0.503	0.586
	CRSAL	0.218	0.390	0.803	0.604	0.719

Numbers in bold-face are the best results for corresponding metrics with p -value < 0.01 .

4.5 Analysis on Effectiveness Results

We first evaluate all models' overall performance on conversational recommendation. The experimental results on three datasets are shown in Table 1.

As illustrated in Table 1, CRSAL outperforms all baselines in terms of the EMR, SR, and QoCR on CamRest676 and DSTC2. Also, CRSAL shows comparable effectiveness in terms of pure dialogue generation quality (BLEU and ROUGE). With the help of the proposed variational inference network and adversarial reinforcement learning method, the NPA in our model can fully discover the dialogue action space with latent action representation and approximate the posterior dialogue action distribution, which means it could generate human-like system actions in each conversation turn to guide the response generation and task completion. Compared with other baselines, CRSAL can effectively improve the recommendation success rate and provide accurate guidance to the NLG to generate a more refined response. We can see that MLEM achieves slightly higher BLEU and ROUGE scores than CRSAL on CamRest676 and DSTC2. This is because it adopts a maximum log-likelihood to directly derive the dialogue action instead of using ELBO to explore the posterior distribution of the action space. It decouples the language generator and the policy agent parameter optimization inside of the variational inference network and gives more attention to the language generation, whereas our proposed model used ELBO to approximate the posterior distribution of the dialogue action and optimized the parameters of the NPA and NLG simultaneously. Thus, with a dedicated loss on action accuracy, the MLEM can achieve relatively higher BLEU scores. However, since BLEU only evaluates how similar the system-generated dialogues are to the human-generated ones, it does not reveal a model's capability of accurately responding to users' demands. Optimizing a conversational recommender system towards BLEU and ROUGE metrics may lead the decoder to overwhelmingly focus on generating better conversations instead of recommending the right items. This explains why we propose QoCR as a joint evaluation metric.

Because in real application scenarios, models that finish the recommendation task with a moderate BLEU loss is more appealing than models that only focus on generating human-like dialogues but fail to provide accurate recommendations.

As presented in Table 1, all the models show decreased performance on MultiWOZ 2.1 dataset in terms of all evaluation metrics. The MultiWOZ 2.1 dataset contains more slot types, slot values, dialogue action types, and linguistic variations in the corpus, which challenge the overall performance of the dialogue agent. Conventional task-oriented dialogue systems highly rely on the explicit human annotations to identify user demands and user intent, which cannot generalize well to complicated conversation corpus like MultiWOZ 2.1. Though being effective on smaller and simpler datasets DSTC2 and CamRest676, the MLEM, however, underperforms on this dataset. This is mainly because MLEM directly uses the log-likelihood to simulate the distribution of the pre-defined dialogue actions. As a result, it performs well when the dialogue action space is limited, as it could lay more emphasis response generation at the same time. However, the MultiWOZ 2.1 dataset contains substantially more dialogue actions than CamRest676 and DSTC2, including multiple types of dialogue actions appearing in one utterance. Therefore, MLEM fails to maintain high accuracy when generating system actions, thus hurting its overall performance. In contrast, CRSAL significantly outperforms all baselines with a large margin on this complex dataset. The ELBO and latent action representations in CRSAL grant the model the ability to fully discover the dialogue action space and generate suitable dialogue actions during the conversation. However, the highly diverse linguistic representations bring challenges to our proposed neural dialogue belief tracker, leading to a lower EMR score compared with that on CamRest676. Furthermore, even though most of the methods achieve a higher EMR score on MultiWOZ 2.1, their task completion (SR) rates are still relatively low. This is because MultiWOZ 2.1 contains more slots, such as time for reservation, number of guests, and so on, and the dialogue state tracking methods in the baselines are insensitive to these slots' numerical variables, which finally causes more failures of the conversation.

Attention should also be paid to the inferior performance of CRS. CRS is designed based on a traditional task-oriented dialogue system, which means its policy agent cannot generate accurate dialogue actions to guide the response generation. In this way, the dialogue agent cannot generate satisfying sentence to interact with the user and identify the user dynamic demand, which finally result in the failure on task-completion and item retrieval, even though it used a recommender system to retrieval the item that meets the user requirements. The LIDM shares the common method with our proposed system to optimize the policy agent. However, it selects the policy gradient-based reinforcement learning to direct fine-tune the parameters of the policy agent without learning the intrinsic evaluation (i.e., the similarity score of the generated response and the reference response) of the generated response. Therefore, our policy agent performs better on action decision and behaves in a more human-like manner.

Another observation is that most of the methods yield the best performance on CamRest676 compared with the other two datasets. This is because CamRest676 is in a relatively small scale and has limited diversity of items. However, CRSAL can still achieve promising results on other two datasets compared with other baseline methods. Last, it can be noticed that the TSCP and LIDM also achieve competitive overall performance on three datasets. As a task-oriented dialogue system, TSCP is relatively robust on conversational recommendation tasks, since the span representation and copynet used in TSCP have greatly helped it with state tracking and response generation in noisy environments. As a state-of-the-art model in latent action modelling, LIDM shows similar effectiveness to TSCP on conversational recommendation tasks. The bottleneck of these two methods is that they cannot learn the intrinsic rewards that can reflect the quality of the generated system actions, which is addressed by the A3C training scheme in our proposed CRSAL model.

Table 2. Ablation Test Results

Dataset	Architecture	BLEU	ROGUE	EMR	SR	QoCR
DSTC2	CRSAL+DDST	0.223	0.384	0.719	0.633	0.745
	CRSAL/A3C	0.212	0.350	0.695	0.605	0.711
	CRSAL+PGRL	0.218	0.374	0.702	0.622	0.735
	CRSAL+A2C	0.220	0.378	0.708	0.622	0.732
	CRSAL+DIR	0.230	0.395	0.712	0.623	0.738
	CRSAL+NDM	0.216	0.362	0.708	0.619	0.726
	Default	0.230	0.395	0.719	0.628	0.743
CamRest676	CRSAL+DDST	0.253	0.478	0.925	0.856	0.983
	CRSAL/A3C	0.221	0.353	0.890	0.735	0.836
	CRSAL+PGRL	0.236	0.398	0.908	0.815	0.933
	CRSAL+A2C	0.238	0.404	0.908	0.821	0.940
	CRSAL+DIR	0.262	0.488	0.915	0.848	0.979
	CRSAL+NDM	0.232	0.391	0.910	0.817	0.933
	Default	0.262	0.488	0.925	0.854	0.985
MultiWOZ 2.1	CRSAL+DDST	–	–	–	–	–
	CRSAL/A3C	0.171	0.298	0.711	0.506	0.592
	CRSAL+PGRL	0.210	0.365	0.779	0.575	0.680
	CRSAL+A2C	0.204	0.343	0.745	0.557	0.659
	CRSAL+DIR	0.218	0.390	0.765	0.573	0.682
	CRSAL+NDM	0.192	0.331	0.730	0.528	0.624
	Default	0.218	0.390	0.803	0.604	0.719

4.6 Model Component Study

Since CRSAL consists of different modules, we analyze the performance gain from different proposed components via a series of ablation experiments. Table 2 presents the performance of our default model architecture and different variants. First, we introduce the variants of CRSAL below:

CRSAL with delexicalization-based dialogue state tracking (CRSAL+DDST): We replace the pre-trained, word vector-based DST in CRSAL with a delexicalization-based DST method to track the dialogue belief state and keep the other components unchanged. The delexicalization-based DST usually requires huge labor to annotate the variant of the lexical representation and link the rephrasings to their ontology. We establish a mapping dictionary to complete this task with the token in the training corpus. We aim to validate that our fully statistical semantic-based DST can ensure the conversational recommendation performance while bypassing the time-consuming, manually crafted semantic annotation dictionaries. It is worth mentioning that we cannot implement CRSAL+DDST on MultiWOZ 2.1, because it is infeasible to manually annotate all the utterances, given its scale and diversity.

CRSAL without A3C (CRSAL/A3C): We remove the entire A3C reinforcement learning module in CRSAL and test the model performance without RL. We want to verify the contribution of the proposed A3C RL method to CRSAL. This degraded model is still trained in a semi-supervised fashion.

CRSAL without A3C, with policy gradient-based RL (CRSAL+PGRL): We use a policy gradient-based (PG) RL method as in Reference [82] to replace our proposed A3C approach to see how our proposed A3C performs in comparison to traditional RL methods.

CRSAL without A3C, with A2C (CRSAL+A2C): We replace the A3C reinforcement learning module with the standard advantage actor-critic algorithm to verify the improvement of our proposed A3C over the A2C-based method.

CRSAL without NRN, with direct item retrieval (CRSAL+DIR): We convert the NRN within CRSAL into a straightforward item retrieval scheme that only uses the extracted dialogue state. We want to see the comparison between our neural recommender network-based system and the retrieval-based system.

CRSAL without NPA and NLG, with neural network (CRSAL+NDM): We replace the NPA and NLG with a simplified network-based dialogue model to demonstrate the performance gain on proposed conversational response generation module.

As shown in Table 2, the proposed adversarial advantage Actor-Critic RL method could effectively enhance the dialogue action modelling. The regulated policy agent shows its strength in improving the task success rate and generating semantic-based responses, which are reflected in higher SR and BLEU scores. The pre-trained word vectors used in our DST could effectively replace the manually crafted semantic matching dictionary and achieve a competitive performance, which means the pre-trained semantic word vectors can eliminate the labor-intensive process in building semantic matching dictionaries. Compared with the conventional policy gradient-based reinforcement learning method, our proposed method benefits from the intrinsic reward from the discriminator and achieves an obvious performance improvement. Also, the baseline function designed in our system could prevent the overfitting of the base model and results in a better performance on all four metrics. From the differences between the retrieval-based methods (CRSAL+DIR) and the default version of CRSAL, we can observe that the CRSAL derives a higher EMR score, because the NRN in CRSAL is prone to recommend the items that are also selected by other users, which is beneficial in cold-start scenarios. Hence, CRSAL can always recommend preferred items to users on both datasets, which makes it outperform the retrieval-based method on EMR.

The experimental results on MultiWOZ 2.1 further demonstrate the superiority of our proposed modules. Due to the complexity of the MultiWOZ 2.1 dataset, it is impractical to construct the delexicalization dictionary linking the lexicon variants of the item feature ontology, so we are unable to test the performance of CRSAL+DDST, but the tests on CamRest676 and DSTC2 have already proved that the proposed NBT in CRSAL is highly competitive in terms of state tracking accuracy. As shown in Table 2, we can observe that the proposed A3C is effective in optimizing the overall performance of CRSAL. Compared with CRSAL+NDM, the proposed NRN in the default version of CRSAL achieves better recommendation accuracy. Particularly, on the largest and most complex dataset MultiWOZ 2.1, it is more advantageous in terms of the EMR metric. The reason behind that is the user usually only gives a small portion of the restaurant attributes to clarify her/his demands in the training corpus, which makes it hard for a straightforward retrieval scheme to infer the right item from limited information. Due to higher expressiveness of the NRN, CRSAL can still yield high recommendation accuracy on large-scale datasets.

Additionally, we conduct further experiments to verify the performance of NBT, because dialogue state tracking is one of the key functions in our proposed model. We use DSTC2 and CamRest676 for this test, as they are well-established benchmark datasets for dialogue state tracking. For a fair comparison, we run our proposed fully statistical NBT and two other popular peer methods for state tracking: one is a delexicalization-based recurrent neural network from Henderson et al. [25] and the other is a data-driven dialogue state tracker proposed by Mrksic et al. [54]. We adopt two common evaluation metrics, namely, Goals and Requests [25, 54]. Specifically, Goals (a.k.a. joint goal accuracy) is the proportion of dialogue turns where all users' search goal constraints are correctly identified, while Requests is the proportion of dialogue turns where users'

Table 3. Dialogue State Tracking Results

	DSTC2		CamRest676	
	Goals	Requests	Goals	Requests
Delexicalization-based Model by Henderson et al. [25]	73.9	96.8	85.7	94.9
Data-driven Model by Mrksic et al. [54]	73.4	96.5	84.8	95.2
Our proposed fully statistical NBT	73.7	96.8	85.9	95.4

requests for information are accurately recognized. From the results presented in Table 3, we can see that our proposed method performs better than the data-driven model in Reference [54]. Compared with Reference [54], our fully statistical NBT introduces a more comprehensive way to integrate the previous dialogue belief state when updating the current one, which improves the overall state tracking performance. Also, it can be observed that our proposed method achieves a similar tracking performance compared with Reference [25], which requires manually crafted semantic dictionaries. It demonstrates that our proposed semantic-based NBT using word embeddings can effectively relieve human annotators from constructing semantic dictionaries for task-oriented dialogue systems.

4.7 Hyperparameter Analysis

The performance of CRSAL is highly correlated with its hyperparameters. Three hyperparameters, namely, the latent action space size H , the weight of semi-supervised loss η , and KL divergence coefficient λ in our system, have deterministic influence on the model performance. In this section, we analyze the model performance w.r.t. different hyperparameter settings. We choose CamRest676 dataset for demonstration due to its high quality on linguistic accuracy, and it is more efficient to train the model with an adequate amount of dialogue actions. Therefore, we conduct hyperparameter analysis on the CamRest676 to showcase the performance fluctuations of CRSAL with different settings. Four evaluation metrics, i.e., EMR, BLEU, SR, and QoCR, are used to illustrate the impact of different hyperparameters.

The impact of latent action space size H is shown in Figure 8. We vary the value of H in $\{50, 70, 100\}$ to observe the system performance change. From Figure 8, we can witness the rise of each metric when H increases from 50 to 70. This is because a larger latent action space could cover more dialogue actions, so a larger proportion of the training set is allocated to training the policy agent in our semi-supervised learning scheme. A fine-grained policy agent could benefit the whole system to generate more accurate response actions and natural language response. The user could clearly express her/his intent and easily communicate with CRSAL, which is reflected by the increase on all four metrics. However, when H increases to 100, it brings a heavy burden for learning a high-dimensional latent action distribution. Our latent action clustering is based on the semantic tokens existing in the user utterances, and if the H is too large, the latent action space will include many irrelevant tokens, which eventually introduces excessive noise for training the policy agent. It has a direct impact on the response generation (BLEU) and also weakens the SR and the joint metric QoCR. However, the EMR metric is not influenced by this change. This reveals that when the policy agent can imitate humans to generate suitable responses, the system could always identify the user intent by the DST and recommend suitable items to the user. So, the EMR does not experience obvious decrease when H increases from 70 to 100.

We also present the influence of the trade-off coefficient η via Figure 9. We observe the model performance with the trade-off coefficient $\eta \in \{0.05, 0.1, 0.3, 0.5, 0.7\}$, while the KL parameter λ remains unchanged as 0.1. From Figure 9, we can observe an overall trend showing a decreasing SR but an improved BLEU score with the increase of η . η is the key factor to balance the importance of

Fig. 8. System performance w.r.t. different latent action space size H . Fig. 9. System performance w.r.t. different trade-off coefficient η .

Fig. 10. System performance w.r.t. different KL parameter λ .

supervised loss and unsupervised loss in our semi-supervised learning scheme. When η is too small (0.05), the system cannot learn sufficient information from the referenced human response and therefore underperforms in terms of task completion and generated response quality, evidenced by lower SR and BLEU scores. With $\eta = 0.1$, the policy agent in CRSAL can learn sufficiently more linguistic information with the unsupervised part. Hence, both SR and BLEU have increased and, consequently, contribute to the increase of QoCR. The improvement of EMR further illustrates that increasing η helps CRSAL capture more accurate information for helping users retrieve their desired restaurants. As η increases from 0.3 to 0.7, the overly strong discriminative power of the language generation model will encourage the whole system to lay more emphasis on generating human-like responses instead of completing the recommendation task, and this is reflected by a gradually growing BLEU score and a significant drop in SR score. Meanwhile, the sub-optimal policy agent weakens the system's ability to respond to user utterances with correct actions, so the system eventually fails to collect enough information for item recommendation and tends to show decreased performance on EMR and QoCR. It is worth mentioning that our proposed QoCR metric integrates the language quality measure with the task completion rate together for a comprehensive evaluation. Based on the saddle point of QoCR, we set $\eta = 0.1$ in our system to ensure sensible model performance on both dialogue generation and item recommendation.

The KL parameter λ in CRSAL is mainly used for optimizing the lower bound of the joint probability of the NLG. It adjusts the weights of the inference network and the KL divergence. We observe the model performance with the KL parameter $\lambda \in \{0.05, 0.1, 0.3, 0.5, 0.7\}$, while the trade-off coefficient η is constantly set to 0.1. Experimental results are demonstrated in Figure 10. We can observe that the BLEU, SR, and QoCR are relatively low with $\lambda = 0.05$. This is because a small λ poses a low weight on the learned information from the language model and weakens the overall performance. As the λ increases, the language model could generate better conversational

responses, and in turn increases the BLEU score. However, the overly strong discriminate power of the language model will disconnect the language model with the rest components in CRSAL and makes the policy agent unable to generate suitable response actions, thus leading to the decrease of SR and QoCR. Also, when the model is laying too much emphasis on forcing the generated system actions to be the same as human responses, the system will consequently lose the ability to adaptively generate further questions to guide the user to clarify her/his intents. As a result, the recommendation effectiveness drops with a large λ , which is reflected by a decreasing EMR score.

4.8 Case Study on Conversational Recommendation

We further investigate the performance of CRSAL by visualizing the real conversation process via Table 4. We compare CRSAL with LIDM, MLEM, and TSCP regarding the quality of their generated responses in the same conversation scenario. The comparison involves the generated responses from different methods as well as the human-made responses, which is referred to as the ground truth (GT).

From the dialogue samples, we can tell many differences between CRSAL and other methods. First, though the four restaurants recommended by other baselines can match this user's description to some extent, only CRSAL recommends the same restaurant as the ground truth to the user. This verifies CRSAL can accurately identify a user's current intent, thus generating the most suitable recommendation results. Then, from the visualized recommendation process, it is clear that CRSAL could generate natural and informative responses to the user utterance, which have successfully guided the user to clarify her/his demand and helped the system to answer the user's additional inquiries. This is because our adversarial Actor-Critic reinforcement learning module could fine-tune our policy agent (NPA) turn-by-turn and give intermediate feedback to optimize the policy agent. In this way, the language generation model NLG also benefits from the NPA in generating high-quality responses.

Besides, although our system outperforms the other baselines on these three datasets, there are still some cases where CRSAL fails to finish the designed task through conversations. We hereby visualize and analyze the unsuccessful samples from CRSAL to seek possible improvements in our future work. Generally, there are two reasons causing the failure of our proposed system during the testing. The first one is that some ground truth sample is unsuccessful as shown in Table 5. To guarantee that the corpus can fully simulate the real human behavior, these three datasets are collected based on interactions between the real humans. Participants are invited to conduct a conversation according to a given task. Most of the given tasks are presented in a way where users first use some attributes to describe their demands and then ask for more details about the recommended item. Few participants would like to finish the task more straightforwardly, so they solely speak out their demands and find the item that can meet their demand, hence they have not finished all the given tasks. As widely acknowledged in the current task-oriented dialogue system research, such incomplete conversations violate the criteria of a successful conversation. The task success rate (SR) defined in Section 4.3 is meant to evaluate a model's ability to generate suitable responses to help users find their desired items and answer the users' queries. If the users in the original corpus do not reach the end of the entire conversational task, then CRSAL cannot acquire further responses from the user, thus being unable to finish the whole task. As presented by Table 5, the given task is to find a restaurant that serves international food in any area and further request some more features of the restaurant. However, the human participant did not express his preference on the price range and failed to select a restaurant that can meet his demand, which is the reason causing the failure of the origin conversation and our proposed CRSAL.

The other reason that causes failure of CRSAL is that, for each conversation, the given task only provides one or two attributes of the restaurant to the conversational recommender

Table 4. An Example of All Models' Conversational Interactions with the Real User

User	I need to find an expensive restaurant that is in the south section of the city.
GT	There are several restaurants in the south part of town that serve expensive food. Do you have a cuisine preference?
CRSAL	There are four restaurants in the south section of the city. Do you have a preference of food type?
LIDM	There are four restaurants in the south section of the city. What kind of food would you like?
MLEM	What food are you looking for?
TSCP	There are four restaurants in the south section of the city. Do you have a preference for the type of food?
User	No, I don't care about the type of cuisine.
GT	Chiquito Restaurant Bar is a Mexican restaurant located in the south part of town.
CRSAL	Chiquito Restaurant Bar is a Mexican restaurant in the south area. Would you like more information?
LIDM	Frankie & Benny's is a Italian restaurant in the south area. Would you like more information?
MLEM	There is the Good Luck. Would you like the information?
TSCP	The Good Luck Chinese food takeaway is a Chinese restaurant in the south area. Would you like their phone number?
User	What is their address ?
GT	Their address is 2G Cambridge Leisure Park Cherry Hinton Road Cherry Hinton, is there anything else I can help you with?
CRSAL	Chiquito Restaurant Bar is located at 2G Cambridge Leisure Park Cherry Hinton Road Cherry Hinton. Is there anything else I can help you with?
LIDM	The address is Cambridge Leisure Park Clifton Way Cherry Hinton.
MLEM	Their address is 82 Cherry Hinton Road Cherry Hinton, is there anything else I can help you with?
TSCP	The Good Luck Chinese food takeaway is located at 82 Cherry Hinton Road Hinton.
User	What is the food type they serve?
GT	They serve Mexican food.
CRSAL	They serve Mexican food.
LIDM	They serve Italian food type.
MLEM	They serve Chinese food.
TSCP	The Good Luck Chinese food takeaway serves Chinese food.
User	OK, that is what I was looking for. Thanks a lot.
GT	Thank you very much for using the Cambridge Restaurant System.
CRSAL	You are welcome. Have a nice day!
LIDM	Have a nice day!
MLEM	Goodbye.
TSCP	Thank you for using the Cambridge Restaurant System. Goodbye.

system retrieving the desired restaurant. In some cases, there would be more than one restaurant that can satisfy the users' demand, which brings challenges to our conversational recommender system. The limited training samples and feature sparsity problem inevitably result in the prediction error on the final recommendation, leading to the failure of this case. From Table 6, we can find that both Nirala and Golden Wok could satisfy the user's demand on price and location. The

Table 5. CRSAL Fails to Provide Further Information Due to Lack of User Response

Task	You want to find a restaurant in any area and it should serve international food. You want to know the address and phone number.
User	I'm looking for a restaurant that serves international food.
GT	There are several restaurants serving international food. Would you prefer one in the center or east part of town?
CRSAL	There are several restaurants serving international food. Would you like to narrow by town?
User	I don't care.
GT	There are three restaurants serving international food. What price range would you like?
CRSAL	There are three restaurants serving international food. Do you have a price range preference?

Table 6. CRSAL Fails to Recommend the Ground Truth Item Due to a Similar Item

Task	You want to find a moderately priced restaurant and it should be in the north part of town. You want to know the address and phone number of the venue.
User	I am looking for a moderately priced restaurant in the north part of town.
GT	The Nirala is located in the north part of town and in the moderate price range.
CRSAL	The Golden Wok is a Chinese restaurant in the north part of the town in the moderate price range.
User	What kind of food is that? Also what is the address and phone number?
GT	Indian food. The address is 7 Milton Road Chesterton and the phone number is 01223 360966. Is there anything else I can help you find?
CRSAL	Chinese food. The address is 191 Histon Road Chesterton and the phone number is 01223 350688. Is there anything I can help you with?
User	Thank you that is all.
GT	You are welcome. Goodbye
CRSAL	You are welcome. Goodbye

difference between these two restaurants is that Nirala provides Indian food, while Golden Wok serves Chinese food. CRSAL finally recommends Golden Wok to the user based on the item's popularity in the training set and closer content information, which causes the failure of the conversation. This problem implies that the model should ask the user to give more details to offer finer information about their demands and preferences, and eventually improve the accuracy of the final recommendation.

4.9 Human Evaluation

To assess the true feedback from the real human users, we additionally employed 30 workers from the crowdsourcing platform Amazon Mechanical Turk to interact with CRSAL and the baselines to qualitatively evaluate the real-life practicality of different conversational recommenders. Each participant is asked to follow a given task and carry out a conversation with the system. At the end of each conversation, each participant is required to evaluate the model performance by giving a score of $\{1, 2, 3, 4, 5\}$ (the higher, the better). The ratings will be given based on the user's experience from the following three aspects. (1) Completion: Has the system successfully finished the

Table 7. Human Evaluation Results

	CRSAL	LIDM	MLEM	TSCP
Completion	4.68	4.60	4.52	4.57
Comprehension	4.37	4.27	3.97	4.15
Naturalness	4.18	4.12	3.95	4.08

recommendation task? (2) Comprehension: Can the system understand the user's queries about the recommended item and respond accordingly? (3) Naturalness: Are the conversations generated by the system human-like? For each model, we collect 200 conversation samples, and the averaged human evaluation scores are shown in Table 7.

We can observe that the human evaluation results show a similar trend as the automatic evaluation results in Table 1. Our proposed CRSAL could outperform all the baselines on all evaluation perspectives with a better generation quality. CRSAL only achieves a marginal gain on the response naturalness, because the semantic-conditioned language generator usually generates the next token conditioned on the sequential semantic information. For completing the utterance semantic representation, the generator sometimes inserts some unnecessary tokens in the generated response, whereas most of the human participants think that tokens should be omitted for better readability, which affects participants' experience with the conversation naturalness. An interesting phenomenon could be observed that the MLEM has obtained the lowest score on comprehension and naturalness. The reason is that the real users' behaviors are more diverse than the ones contained in the training corpus. The maximum likelihood action estimation in MLEM cannot support high quality-language interpretation when the conversation scenario becomes complex, which finally results in the collapse of human evaluation performance.

5 CONCLUSION

In this article, we have proposed CRSAL, a new conversational recommender system, which integrates the dialogue system with a recommender system seamlessly to perform personalized recommendation via human-like conversation. Distinguished from the conventional recommender system, our proposed CRSAL improves the user experience with a more accessible recommendation method by conversation. It not only helps customers clarify their demands and find ideal items, but also answers users' extended queries about the recommendations. CRSAL additionally tackles the drawbacks of existing task-oriented dialogue systems for conversational recommendation with a novel adversarial reinforcement learning framework to achieve better performance.

The whole model is trained in a semi-supervised fashion. It mainly consists of three key modules: information extraction, conversational response generation, and adversarial reinforcement learning. The information extraction module analyzes the input user utterance to track the current dialogue state, extract the users' intent, and generate a personalized recommendation list. The output of the information extraction is fed into conversational response generation to help to train a neural policy agent and a natural language generator via a variational inference network. An adversarial advantage Actor-Critic reinforcement learning method is finally adopted to encourage the policy agent to generate human-like actions, thus improving the overall system preference. We verified our proposed method on three public datasets and evaluated it via four widely accepted metrics, i.e., BLEU, ROGUE, entity match rate, and task success rate. We also propose a new metric QoCR for conversational recommendation via a trade-off between the language quality and task success rate. Experimental results verify that the proposed method could achieve a better performance compared with state-of-the-art methods. Additional ablation tests and parameter tests further study the internal performance of our proposed method. Our method

could obtain a considerable performance compared with methods that rely on manually crafted resources.

One interesting and meaningful further direction is to extend our proposed system to multi-domain scenarios. The user sometimes will want results from various domains instead of a single one to meet her/his diverse demands by one single system. However, as pointed out by References [5, 98], the different conversation domain may involve different conversation actions, which brings the challenge of dialogue action representation learning. To this end, we will explore the latent action representations in further work and utilize them to further improve the quality of generated responses in a multi-domain conversational recommendation scenario.

ACKNOWLEDGMENTS

We would like to thank our anonymous reviewers for providing insightful review comments and suggestions.

REFERENCES

- [1] J. E. Allen, C. I. Guinn, and E. Horvitz. 1999. Mixed-initiative interaction. *IEEE Intell. Syst. Their Applic.* 14, 5 (1999), 14–23. DOI: <https://doi.org/10.1109/5254.796083>
- [2] Suman Banerjee and Mitesh M. Khapra. 2019. Graph convolutional network with sequential attention for goal-oriented dialogue systems. *Trans. Assoc. Comput. Ling.* 7 (2019), 485–500. DOI: https://doi.org/10.1162/tacl_a_00284
- [3] Antoine Bordes and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of the International Conference on Learning Representations (ICLR'17)*. Retrieved from <http://arxiv.org/abs/1605.07683>.
- [4] Joshua W. Buck, Saverio Perugini, and Tam V. Nguyen. 2018. Natural language, mixed-initiative personal assistant agents. In *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication (IMCOM'18)*. DOI: <https://doi.org/10.1145/3164541.3164609>
- [5] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ—A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*. Association for Computational Linguistics, 5016–5026. Retrieved from <https://www.aclweb.org/anthology/D18-1547>.
- [6] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explor. Newslett.* 19, 2 (2017), 25–35.
- [7] Q. Chen and W. Wang. 2019. Sequential matching model for end-to-end multi-turn response selection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*. 7350–7354. DOI: <https://doi.org/10.1109/ICASSP.2019.8682538>
- [8] Tianqi Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'18)*. 2610–2620. Retrieved from <http://papers.nips.cc/paper/7527-isolating-sources-of-disentanglement-in-variational-autoencoders.pdf>.
- [9] Tong Chen, Hongzhi Yin, Hongxu Chen, Rui Yan, Quoc Hung Nguyen, and Xue Li. 2019. AIR: Attentional intention-aware recommender systems. In *Proceedings of the IEEE 35th International Conference on Data Engineering (ICDE'19)*. 304–315.
- [10] Tong Chen, Hongzhi Yin, Quoc Viet Hung Nguyen, Wen-Chih Peng, Xue Li, and Xiaofang Zhou. 2020. Sequence-aware factorization machines for temporal predictive analytics. In *Proceedings of the IEEE 36th International Conference on Data Engineering (ICDE'20)*. 1405–1416.
- [11] Y. Chen, W. Y. Wang, and A. I. Rudnicky. 2013. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*. 120–125. DOI: <https://doi.org/10.1109/ASRU.2013.6707716>
- [12] Zhiyong Cheng and Jialie Shen. 2016. On effective location-aware music recommendation. *ACM Trans. Inf. Syst.* 34, 2, Article 13 (Apr. 2016), 13:1–13:32 pages. DOI: <https://doi.org/10.1145/2846092>
- [13] 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 and Data Mining (KDD'18)*. ACM, 139–148. DOI: <https://doi.org/10.1145/3219819.3219894>
- [14] Gregorio Convertino, Massimo Zancanaro, Tiziano Piccardi, and Felipe Ortega. 2017. Toward a mixed-initiative QA system: From studying predictors in stock exchange to building a mixed-initiative tool. *Int. J. Hum.-Comput. Stud.* 99 (2017), 20. 1–20. DOI: <https://doi.org/10.1016/j.jhcs.2016.10.008>

- [15] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-N recommendation tasks. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10)*. ACM, 39–46. DOI: <https://doi.org/10.1145/1864708.1864721>
- [16] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperAgent: A customer service chatbot for e-commerce websites. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (ACL'17)*. Association for Computational Linguistics, 97–102. Retrieved from <https://www.aclweb.org/anthology/P17-4017>.
- [17] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (ACL'17)*. Association for Computational Linguistics, 484–495. DOI: <https://doi.org/10.18653/v1/P17-1045>
- [18] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. 2019. MultiWOZ 2.1: Multi-domain dialogue state corrections and state tracking baselines. *CoRR* abs/1907.01669 (2019).
- [19] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational AI. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)*. ACM, 1371–1374. DOI: <https://doi.org/10.1145/3209978.3210183>
- [20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'14)*. Curran Associates, Inc., 2672–2680. Retrieved from <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>.
- [21] Lei Guo, Hongzhi Yin, Qinyong Wang, Tong Chen, Alexander Zhou, and Nguyen Quoc Viet Hung. 2019. Streaming session-based recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'19)*. ACM, 1569–1577. DOI: <https://doi.org/10.1145/3292500.3330839>
- [22] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, Vol. 80. PMLR, 1861–1870. Retrieved from <http://proceedings.mlr.press/v80/haarnoja18b.html>.
- [23] Xiangnan He and Tat-Seng Chua. 2017. Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'17)*. ACM, 355–364. DOI: <https://doi.org/10.1145/3077136.3080777>
- [24] Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'14)*. Association for Computational Linguistics, 263–272. DOI: <https://doi.org/10.3115/v1/W14-4337>
- [25] M. Henderson, B. Thomson, and S. Young. 2014. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT'14)*. 360–365. DOI: <https://doi.org/10.1109/SLT.2014.7078601>
- [26] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR'16)*.
- [27] Eric J. Horvitz. 2007. Reflections on challenges and promises of mixed-initiative interaction. *AI Mag.* 28, 2 (June 2007), 19–22. DOI: <https://doi.org/10.1609/aimag.v28i2.2036>
- [28] Rein Houthoofd, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2016. VIME: Variational information maximizing exploration. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'16)*. 1109–1117. Retrieved from <http://papers.nips.cc/paper/6591-vime-variational-information-maximizing-exploration.pdf>.
- [29] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. 2017. Collaborative metric learning. In *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*.
- [30] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the IEEE International Conference on Data Mining*.
- [31] Neil J. Hurley. 2013. Personalised ranking with diversity. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys'13)*. ACM, 379–382. DOI: <https://doi.org/10.1145/2507157.2507226>
- [32] Xisen Jin, Wenqiang Lei, Zhaochun Ren, Hongshen Chen, Shangsong Liang, Yihong Zhao, and Dawei Yin. 2018. Explicit state tracking with semi-supervision for neural dialogue generation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18)*. ACM, 1403–1412.
- [33] J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.* 2, 1 (Jan. 1984), 26–41. DOI: <https://doi.org/10.1145/357417.357420>

- [34] Yoon Kim, Sam Wiseman, Andrew Miller, David Sontag, and Alexander Rush. 2018. Semi-amortized variational autoencoders. In *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, Vol. 80. PMLR, 2678–2687. Retrieved from <http://proceedings.mlr.press/v80/kim18e.html>.
- [35] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR'15)*.
- [36] Yehuda Koren. 2008. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, 426–434. DOI : <https://doi.org/10.1145/1401890.1401944>
- [37] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics (ACL'18)*. Association for Computational Linguistics, 1437–1447. Retrieved from <https://www.aclweb.org/anthology/P18-1133>.
- [38] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*.
- [39] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM'17)*. ACM, 1419–1428. DOI : <https://doi.org/10.1145/3132847.3132926>
- [40] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'18)*. Curran Associates, Inc., 9725–9735. Retrieved from <http://papers.nips.cc/paper/8180-towards-deep-conversational-recommendations.pdf>.
- [41] Xinyi Li, Yifan Chen, Benjamin Pettit, and Maarten De Rijke. 2019. Personalised reranking of paper recommendations using paper content and user behavior. *ACM Trans. Inf. Syst.* 37, 3, Article 31 (Mar. 2019), 31:1–31:23 pages. DOI : <https://doi.org/10.1145/3312528>
- [42] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81. Retrieved from <https://www.aclweb.org/anthology/W04-1013>.
- [43] Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. 2018. BBQ-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 5237–5244. Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16189/16073>.
- [44] Bing Liu and Ian Lane. 2018. Adversarial learning of task oriented neural dialog models. In *Proceedings of the 19th SIGdial Meeting on Discourse and Dialogue*. 147–157.
- [45] Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'18)*. Association for Computational Linguistics, 2060–2069. DOI : <https://doi.org/10.18653/v1/N18-1187>
- [46] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2122–2132. DOI : <https://doi.org/10.18653/v1/D16-1230>
- [47] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: Short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18)*. ACM, 1831–1839. DOI : <https://doi.org/10.1145/3219819.3219950>
- [48] Chris J. Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. 2017. Filtering variational objectives. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. Curran Associates, Inc., 6573–6583. Retrieved from <http://papers.nips.cc/paper/7235-filtering-variational-objectives.pdf>.
- [49] Prem Melville and Vikas Sindhwani. 2017. Recommender systems. *Encyclopedia of Machine Learning and Data Mining*. Springer.
- [50] Yishu Miao and Phil Blunsom. 2016. Language as a latent variable: Discrete generative models for sentence compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*. Association for Computational Linguistics, 319–328. DOI : <https://doi.org/10.18653/v1/D16-1031>
- [51] Volodymyr Mnih, Adria Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of*

- the 33rd International Conference on International Conference on Machine Learning (ICML'16). JMLR.org, 1928–1937. Retrieved from <http://dl.acm.org/citation.cfm?id=3045390.3045594>.
- [52] Nikola Mrkšić, Diarmuid Ó. Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'16)*. Association for Computational Linguistics, 142–148. DOI : <https://doi.org/10.18653/v1/N16-1018>
- [53] Nikola Mrkšić, Diarmuid Ó. Séaghdha, Blaise Thomson, Milica Gašić, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain dialog state tracking using recurrent neural networks. In *Proceedings of the 53rd Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL'15)*.
- [54] Nikola Mrksic, Diarmuid Ó. Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (ACL'17)*. 1777–1788. DOI : <https://doi.org/10.18653/v1/P17-1163>
- [55] Nikola Mrkšić and Ivan Vulić. 2018. Fully statistical neural belief tracking. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics (ACL'18)*. Association for Computational Linguistics, 108–113. Retrieved from <https://www.aclweb.org/anthology/P18-2018>.
- [56] 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 Meeting on Association for Computational Linguistics (ACL'02)*. Association for Computational Linguistics, 311–318. DOI : <https://doi.org/10.3115/1073083.1073135>
- [57] Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Chen, Yun-Nung Liu, and Kam-Fai Wong. 2018. Adversarial advantage actor-critic model for task-completion dialogue policy learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*. 6149–6153.
- [58] Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep Dyna-Q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics (ACL'18)*. Association for Computational Linguistics, 2182–2192. Retrieved from <https://www.aclweb.org/anthology/P18-1203>.
- [59] Bilih Priyogi. 2019. Preference elicitation strategy for conversational recommender system. In *Proceedings of the International Conference on Web Search and Data Mining*.
- [60] S. Rao and Michael Wooldridge. 1999. *Foundations of Rational Agency*. Springer.
- [61] Steffen Rendle. 2010. Factorization machines. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'10)*. 995–1000. DOI : <https://doi.org/10.1109/ICDM.2010.127>
- [62] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09)*. AUAI Press, 452–461. Retrieved from <http://dl.acm.org/citation.cfm?id=1795114.1795167>.
- [63] Steffen Rendle, Zeno Gantner, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2011. Fast context-aware recommendations with factorization machines. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*.
- [64] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML'14)*, Vol. 32. JMLR.org, II–1278–II–1286. Retrieved from <http://dl.acm.org/citation.cfm?id=3044805.3045035>.
- [65] Ruhi Sarikaya, Geoffrey E. Hinton, and Bhuvan Ramabhadran. 2011. Deep belief nets for natural language call-routing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'11)*. 5680–5683. DOI : <https://doi.org/10.1109/ICASSP.2011.5947649>
- [66] Martin Saveski and Amin Mantrach. 2014. Item cold-start recommendations: Learning local collective embeddings. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys'14)*. ACM, 89–96. DOI : <https://doi.org/10.1145/2645710.2645751>
- [67] Lei Shi, Wayne Xin Zhao, and Yi-Dong Shen. 2017. Local representative-based matrix factorization for cold-start recommendation. *ACM Trans. Inf. Syst.* 36, 2, Article 22 (Aug. 2017), 22:1–22:28 pages. DOI : <https://doi.org/10.1145/3108148>
- [68] Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. Continuously learning neural dialogue management. *CoRR abs/1606.02689* (2016).
- [69] Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics (ACL'16)*. 2431–2441. Retrieved from <http://arxiv.org/abs/1605.07669>.

- [70] Pei-Hao Su, Pawel Budzianowski, Stefan Ultes, Milica Gasić, and Steve Young. 2017. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In *Proceedings of the 18th SIGdial Meeting on Discourse and Dialogue (SIGDIAL'17)*.
- [71] Shang-Yu Su, Xiuju Li, Jianfeng Gao, Jingjing Liu, and Yun-Nung Chen. 2018. Discriminative deep Dyna-Q: Robust planning for dialogue policy learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*. Association for Computational Linguistics, 3813–3823. Retrieved from <https://www.aclweb.org/anthology/D18-1416>.
- [72] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'18)*. ACM, 235–244. DOI: <https://doi.org/10.1145/3209978.3210002>
- [73] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'14)*. Curran Associates, Inc., 3104–3112. Retrieved from <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>.
- [74] Qinyong Wang, Hongzhi Yin, Zhiting Hu, Defu Lian, Hao Wang, and Zi Huang. 2018. Neural memory streaming recommender networks with adversarial training. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18)*. ACM, 2467–2475. DOI: <https://doi.org/10.1145/3219819.3220004>
- [75] Qinyong Wang, Hongzhi Yin, Hao Wang, Quoc Viet Hung Nguyen, Zi Huang, and Lizhen Cui. 2019. Enhancing collaborative filtering with generative augmentation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*. 548–556. DOI: <https://doi.org/10.1145/3292500.3330873>
- [76] Weiqing Wang, Hongzhi Yin, Zi Huang, Qinyong Wang, Xingzhong Du, and Quoc Viet Hung Nguyen. 2018. Streaming ranking based recommender systems. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)*. ACM, 525–534. DOI: <https://doi.org/10.1145/3209978.3210016>
- [77] Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based RNN semantic frame parsing model for intent detection and slot filling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'18)*. Association for Computational Linguistics, 309–314. DOI: <https://doi.org/10.18653/v1/N18-2050>
- [78] Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. 2017. Sample efficient actor-critic with experience replay. In *Proceedings of the International Conference on Learning Representations*.
- [79] Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. AirDialogue: An environment for goal-oriented dialogue research. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*. Association for Computational Linguistics, 3844–3854. Retrieved from <https://www.aclweb.org/anthology/D18-1419>.
- [80] Gellert Weisz, Pawel Budzianowski, Pei-Hao Su, and Milica Gasic. 2018. Sample efficient deep reinforcement learning for dialogue systems with large action spaces. *IEEE/ACM Trans. Audio, Speech Lang. Proc.* 26, 11 (Nov. 2018), 2083–2097. DOI: <https://doi.org/10.1109/TASLP.2018.2851664>
- [81] Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*. JMLR.org, 3732–3741. Retrieved from <http://dl.acm.org/citation.cfm?id=3305890.3306067>.
- [82] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*. Association for Computational Linguistics, 438–449. Retrieved from <https://www.aclweb.org/anthology/E17-1042>.
- [83] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Trans. Assoc. Comput. Ling.* 3 (2015), 345–358. DOI: https://doi.org/10.1162/tacl_a_00143
- [84] Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: Practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (ACL'17)*. Association for Computational Linguistics, 665–677. DOI: <https://doi.org/10.18653/v1/P17-1062>
- [85] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J. Smola, and How Jing. 2017. Recurrent recommender networks. In *Proceedings of the International Conference on Web Search and Data Mining*.
- [86] Wei Wu and Rui Yan. 2019. Deep chit-chat: Deep learning for chatbots. In *Proceedings of the World Wide Web Conference (WWW'19)*. ACM, 1329–1329. DOI: <https://doi.org/10.1145/3308560.3320084>
- [87] Wei Wu and Rui Yan. 2019. Deep chit-chat: Deep learning for chatbots. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1413–1414.
- [88] Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the Conference of the North American Chapter of the Association for Computational*

- Linguistics (NAACL'19)*. Association for Computational Linguistics, 2324–2335. Retrieved from <https://www.aclweb.org/anthology/N19-1242>.
- [89] Haotian Xu, Haiyun Peng, Haoran Xie, Erik Cambria, Liuyang Zhou, and Weiguo Zheng. 2019. End-to-End latent-variable task-oriented dialogue system with exact log-likelihood optimization. *World Wide Web* (07 June 2019). DOI: <https://doi.org/10.1007/s11280-019-00688-8>
 - [90] Rui Yan and Dongyan Zhao. 2018. Coupled context modeling for deep chit-chat: Towards conversations between human and computer. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18)*. ACM, 2574–2583. DOI: <https://doi.org/10.1145/3219819.3220045>
 - [91] Hongzhi Yin, Bin Cui, Yizhou Sun, Zhiting Hu, and Ling Chen. 2014. LCARS: A spatial item recommender system. *ACM Trans. Inf. Syst.* 32, 3, Article 11 (2014), 11:1–11:37 pages. DOI: <https://doi.org/10.1145/2629461>
 - [92] Hongzhi Yin, Bin Cui, Xiaofang Zhou, Weiqing Wang, Zi Huang, and Shazia Sadiq. 2016. Joint modeling of user check-in behaviors for real-time point-of-interest recommendation. *ACM Trans. Inf. Syst.* 35, 2 (2016). DOI: <https://doi.org/10.1145/2873055>
 - [93] Hongzhi Yin, Weiqing Wang, Hao Wang, Ling Chen, and Xiaofang Zhou. 2017. Spatial-aware hierarchical collaborative deep learning for POI recommendation. *IEEE Trans. Knowl. Data Eng.* 29, 11 (2017), 2537–2551.
 - [94] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.* 52, 1, Article 5 (Feb. 2019), 38 pages. DOI: <https://doi.org/10.1145/3285029>
 - [95] 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 (CIKM'18)*. ACM, 177–186. DOI: <https://doi.org/10.1145/3269206.3271776>
 - [96] Zheng Zhang, Minlie Huang, Zhongzhou Zhao, Feng Ji, Haiqing Chen, and Xiaoyan Zhu. 2019. Memory-augmented dialogue management for task-oriented dialogue systems. *ACM Trans. Inf. Syst.* 37, 3, Article 34 (July 2019), 34:1–34:30 pages. DOI: <https://doi.org/10.1145/3317612>
 - [97] Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the 17th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'16)*. Association for Computational Linguistics, 1–10. DOI: <https://doi.org/10.18653/v1/W16-3601>
 - [98] Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'19)*. Association for Computational Linguistics, 1208–1218. Retrieved from <https://www.aclweb.org/anthology/N19-1123>.

Received September 2019; revised April 2020; accepted April 2020