A Versatile Adaptive Curriculum Learning Framework for Task-oriented Dialogue Policy Learning

Yangyang Zhao[♣§] , **Hua Qin**[♣] ,**Zhenyu Wang**[♣†], **Changxi Zhu**[§] and **Shihan Wang**[§‡]
[♣]South China University of Technology
[§]Utrecht University
msyyz@mail.scut.edu.cn; qhscut@163.com; wangzy@scut.edu.cn;
cxzhu.cn@gmail.com; s.wang2@uu.nl

Abstract

Training a deep reinforcement learning-based dialogue policy with brute-force random sampling is costly. A new training paradigm was proposed to improve learning performance and efficiency by combining curriculum learning. However, attempts in the field of dialogue policy are very limited due to the lack of reliable evaluation of difficulty scores of dialogue tasks and the high sensitivity to the mode of progression through dialogue tasks. In this paper, we present a novel versatile adaptive curriculum learning (VACL) framework, which presents a substantial step toward applying automatic curriculum learning on dialogue policy tasks. It supports evaluating the difficulty of dialogue tasks only using the learning experiences of dialogue policy and skip-level selection according to their learning needs to maximize the learning efficiency. Moreover, an attractive feature of VACL is the construction of a generic, elastic global curriculum while training a good dialogue policy that could guide different dialogue policy learning without extra effort on re-training. The superiority and versatility of VACL are validated on three public dialogue datasets.

1 Introduction

Deep reinforcement learning (DRL) as a promising method has impressive achievements for dialogue policy learning (Young et al., 2013; Gasic et al., 2013; Su et al., 2018; Zhang et al., 2019; Lubis et al., 2020; Wang et al., 2021; Zhao et al., 2021b). But current DRL-based dialogue policy approaches mostly remain brute-force random sampling training, improving their performance at the expense of high interaction costs (Jiang et al., 2015; Ren et al., 2018; Narvekar and Stone, 2019; Narvekar et al., 2020). Inspired by human education, a novel training paradigm, curriculum learning (CL), is proposed to improve learning performance and efficiency through training a model on a designed sequence of training tasks, rather than an arbitrary random sampling (Svetlik et al., 2017; Weinshall et al., 2018; Fan et al., 2018; Racanière et al., 2019; Green et al., 2019). Although many empirical studies demonstrated beneficial effects of CL, reporting in the field of dialogue policy remains very limited (Zhao et al., 2021a; Liu et al., 2021).

One reason for the slow adoption of such approaches is the absence of a reliable evaluation of the difficulty score of a dialogue task (Weinshall and Amir, 2018). On the other hand, the high sensitivity to the mode of progression through dialogue tasks makes it difficult to adaptively select the dialogue task at the appropriate difficulty for the current dialogue policy (Graves et al., 2017).

In this paper, we propose a versatile adaptive curriculum learning (VACL) framework, which consists of two main components: One is a difficulty measurer that evaluates the difficulty of dialogue tasks only using learning experiences of the dialogue policy (also refer to student model) and ranks them by difficulty to obtain a global curriculum, exempting from the limitation of unable human-defined difficulty. A generic, elastic global curriculum is available after the end of training. The other is a training scheduler that supports skip-level selection on the global curriculum according to the learning needs of the student model to maximize the learning efficiency. The skip-level selection could either choose a harder task to avoid wasting time on too easy tasks or return to an easier task to prevent forgetting. Besides, an attractive feature of VACL is the construction of a generic, elastic global curriculum while training a good dialogue policy that could guide different dialogue policy learning without extra effort on re-training. Our model is model-agnostic, in the sense that it can be incorporated into different student models.

[*]Contribute equally
[†]Corresponding author
[‡]Corresponding author

Findings of the Association for Computational Linguistics: NAACL 2022, pages 711 - 723
July 10-15, 2022

In summary, the main contributions of this paper are three-fold: 1) We propose a novel VACL framework, which presents a substantial step toward applying automatic curriculum learning on dialogue policy tasks. 2) We explore the versatility of VACL that it succeeds in training a good dialogue policy while building a generic curriculum to guide diverse student learning without extra effort on re-training. To our knowledge, it is the first to demonstrate the extra value of curriculum learning in dialogue policy tasks. 3) We validate the superiority and versatility of VACL on three public dialogue datasets. Additionally, we analyzed the ranking of the generic global curriculum, which gave us some inspiration to refine the difficulty criteria for dialogue tasks.

2 Related Work

Inspired by the human education process, it is not new to apply curriculum learning into dialogue policy tasks to restructure the training process of dialogue agents (Selfridge et al., 1985). There are two categories for those approaches, one at the experiences replay level for data exploitation and the other at the task level for data collection (Portelas et al., 2020). This paper focus here on the second kind of study.

The methods at the experience replay level can be considered as a ranking of the transitions, which can be implemented by *transition selection* or *transition modification*. Prioritized experience replay (PER), a typical *transition selection* method, biased selects transitions with higher TD-error (Schaul et al., 2016). However, PER is highly sensitive to parameter changes, too large parameter changes make it difficult for PER to convergence. Hindsight experience replay (HER) controls the distribution of training transitions by creating successful dialogue experiences from failed ones (Lu et al., 2019). However, not all failed experiences are equally useful for improving dialogue agents. Some provide limited help in reaching the complete tasks, while some are too similar to each other and thus redundant to be learned entirely.

In contrast, there are very limited attempts on researches at the task level. Zhao et al. (2021a) presented preliminary attempts, namely Automatic Curriculum Learning-based Deep Q-Network (ACL-DQN). The ACL-DQN achieves teacher-student co-evolution by incorporating a teacher model to leverage the over-repetition rewards and the feedback from the student model to optimize its curriculum by means of RL. But it remains the drawback of RL, that is, training a good teacher model from scratch may require more experience than learning a good student model. Thus, the method equips the RL-based teacher model with different hand-crafted courses to assist it in customizing good curriculums even in the early training phase. However, the equipped courses presuppose user goals[1] can be ordered by the number of slots in user goals, whereas in reality, they may vary along with multiple factors (Kim and Choi, 2018), an observation that has been verified in our experiment. Recently, Liu et al. (2021) has proposed a novel Scheduled Dialog Policy Learning (SDPL) approach to assess the difficulty of user goals using the dialogue state differential space and scale the capacity of the training goal set proportionally as the training time increases. However, such a difficulty measurer is highly affected by the training dynamics of the model itself, making its curriculum only applicable to the model being trained this time. This implies that user goals that were not or less learned typically be more difficult, and retraining may yield different orders. And the pacing function in SDPL ignores the learning feedback of dialogue policies, resulting in forgotten problems and performance impairment.

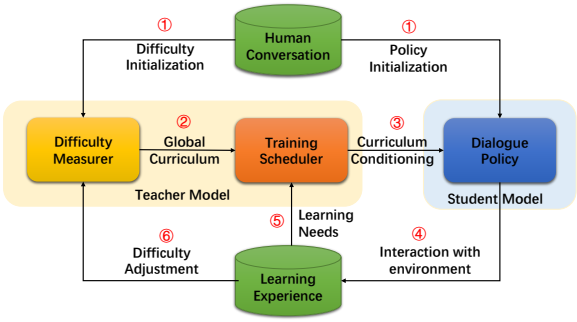


Figure 1: Proposed VACL framework.

3 Proposed Methods

The overall framework of the Versatile Adaptive Curriculum learning is shown in Figure 1, which includes two sub-modules: A *difficulty measurer* that measures the difficulty of user goals and ranks them by difficulty to obtain a global curriculum. After the end of training, a generic, elastic global curricu-

[1]In the field of dialogue policy, each user goal is considered a dialogue task. Readers can refer to Appendix .1 for details on the user goal.

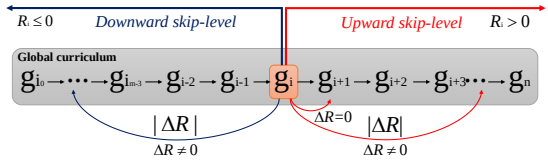


Figure 2: Training scheduler schematic.

lum is available to guide different students learning without extra efforts on re-training; A *training scheduler* model that follows the global curriculum and adaptively curriculum-conditioning according to the learning needs of the student model to maximize learning efficiency. We will describe the two modules in detail and the implementation of the integrated VACL algorithm in what follows.

3.1 Difficulty Measurer

In the context of human education, the distribution of students' exam scores reflects the difficulty of the test. We build on this intuition for constructing a difficulty measurer to evaluate the difficulty of user goals by calculating the average cumulative reward of their related samples. The greater the average cumulative returns, the better students' mastery of this user goal, which means that this user goal is easier for the student. Hence, we define user goal difficulty as:

Definition 1. *(**User Goal Difficulty**): we assume that there exist interaction trajectories for each user goal, $\{s_0, a_0, s_1, a_1, s_2, a_2, ...\}$, whose corresponding return trajectories are $\{r_0, r_1, r_2, ...\}$. We define the average cumulative return of these sample trajectories about this user goals as the difficulty of the user goal g_i, called the user goal value function $V(g)$.*

Average cumulative returns are attractive because they are usually cheaper to obtain and provide a more measured evaluation for user goals' difficulty without knowing the overall user goals. Since user goal value functions $V(g)$ are presumed by sampling return trajectories and average their cumulative returns, we apply the Monte Carlo method (Hammersley, 2013) to evaluate the user goal value functions without manually presuming unknown difficulty factors. Given a user goal g, its user goal value function is evaluated as follows:

$$V(g) = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{M_j} \gamma^k r_k^j \qquad (1)$$

where N denotes the number of all interaction trajectories corresponding to the user goal g, and M_j denotes the length of the sampled interaction trajectories j.

The difficulty measurer ranks user goals from easy to difficult based on calculated user goal difficulties to obtain a global curriculum for the training scheduler. New interaction trajectories generated from student learning are also taken into account in the update of the user goal value function:

$$V_{t+1}(g) = V_t(g) + \alpha(R_{t+1} - V_t(g)) \qquad (2)$$

where α denotes the update rate, and we experimentally evaluate the impact of varying α on performance in Appendix .3. $V_{t+1}(g)$ and $V_t(g)$ denote the value function of the user goal g at the previous and current moments, respectively. R_{t+1} denotes the cumulative return of new interaction trajectory about the user goal g.

As training times increases and more interaction samples are collected, the user goal value functions gradually converge. Through sorting the converging user goal value functions, a generic, elastic global curriculum is obtained, which could be applied to new students learning without extra efforts on re-training [2]. The global curriculum storage form is as follows:

$$[\{g_0 : V(g_0)\}, \{g_1 : V(g_1)\}, [\{g_2 : V(g_2)\}...]$$

where g_i denotes the index of the user goal g_i, and $V(g_i)$ denotes the difficulty of the user goal g_i.

3.2 Training Scheduler

Ideally, students should follow a curriculum based on global difficulty, while not wasting time on tasks that are easy for their current learning ability. As a result, instead of selecting from the global curriculum one by one, the training scheduler in VACL allows skip-level selection on the global curriculum according to the learning needs of the student model to maximize learning efficiency.

To achieve this goal, we first outline the learning needs for the student model as follows: In principle, training on easy user goals gains less information than difficult ones, and direct training on

[2]The relative order of the global curriculum remains roughly the same even for different dialogue policies, despite the change in their difficulty score. For example, for students at any grade level, high school lessons are always harder than elementary schools.

difficult user goals may not gain positive guidance. To maximize access to information and establish the correct location of the decision surface, student models tend to select a user goal with difficulty that matches their ability. Hence, the suitable difficulty of a well-match user goal for the dialogue policy in the current stage is satisfied as follows [3]:

$$D(g') = \begin{cases} V(g_i) - |\Delta R|, & \text{if } R_i > 0 \text{ or } \Delta R = 0, \\ V(g_i) + |\Delta R|, & else \end{cases}$$

$$\Delta R = \begin{cases} R_i - V(g_i), & \text{if } R_i \neq V(g_i), \\ V(g_{i+1}) - V(g_i), & else \end{cases} \quad (3)$$

where $V(g_i)$ denotes the value function of the current user goal g_i, R_i denotes the current cumulative returns obtained when completing the user goal g_i, and ΔR measures the current gap between the ability of the student policy and the user goal difficulty.

As shown in Fig 2, R_i controls the direction of the next user goal selection in the global curriculum, while ΔR controls the distance. $R_i > 0$ indicates that the dialogue policy already successfully completed the user goal g_i, and the next task tends to be more difficult. A high ΔR indicates low learning gains due to a large gap between the capability of dialogue policy and the difficulty of the current user goal, and the next curriculum tends to choose a user goal that is further away from the current user goal in the global curriculum. The training scheduler continues to follow the global curriculum when the capability of student policy and the user goal difficulty are comparable, $\Delta R = 0$.

Finally, the training scheduler selects a user goal with the difficulty closest to the suitable difficulty $D(g')$ for the next training, as shown in Algorithm 1.

3.3 Implementation

Teaching according to aptitude accelerates the learning efficiency of the dialogue policy, and the performance-enhanced dialogue policy generates higher quality sample trajectories, which further improve the accuracy and rationality of the teacher model in difficulty assessing and curriculum arrangement. Both of them promote each other and jointly improve the learning efficiency of dialogue policies. The implementation of the VACL framework is shown in Algorithm 2.

[3]Since the user goal value function and the learning gains belong to the same order of magnitude, it is computationally desirable.

Algorithm 1 Find the next user goal $find\_nearest(curriculum, value)$

Input: The global curriculum S and the suitable difficulty $D(g')$

Output: Index of the well-match user goal for the next training g'

Initialize an empty list A to store all user goal difficulties in S
for $i \leftarrow len(S)$ **do**
  $A.append(S[i].values())$
end for
$A \leftarrow np.array(A)$
$i \leftarrow (np.abs(A - D(g'))).argmin()$
$g' \leftarrow S[i].keys()$
return g'

Algorithm 2 VACL for Dialogue Policy Learning.

Initialize an empty experience relay buffer D and a list T to store user goals index and their difficulty pairs
① Policy and Difficulty Initialized
Initialize $Q(s, a; \theta_Q)$ via pre-training on human conversational data
The difficulty of each user goal $V(g)$ in human conversational data is calculated via Eq. 1 and stored in T together with their corresponding user goal g
Initialize an initial suitable difficulty $D(g) = (V(g)_{max} + V(g)_{min})/2$
for $t = 1 : N$ **do**
  ② Global Curriculum Generation
  The difficulty measurer ranks T by difficulty to obtain a global curriculum S
  ③ Curriculum Conditioning
  The training scheduler selects a user goal g based on the suitable difficulty $D(g)$, $g = find\_nearest(S, D(g))$
  ④ Interaction with environment
  A User and a dialogue agent interact around selected user goal g and their generated experiences are stored in D
  ⑤ Measure Learning needs
  The training scheduler measures the learning needs according to the current cumulative rewards R and evaluates the next suitable difficulty $D(g)$ via Eq. 3
  ⑥ Difficulty Adjustment
  The difficulty measurer also uses R to update the user goal difficulty via Eq. 2
  Sample random mini-batches from D and update θ_Q via M-step minibatch Q-learning
end for

4 Experiments

To evaluate the effectiveness and versatility of the proposed VACL framework, experiments are carried out on three public single datasets, Movie-Ticket Booking, Restaurant Reservation, Taxi Ordering (Li et al., 2016, 2018) in both simulation and human evaluation.

4.1 Dataset

The experiments are carried out on the platform Microsoft Dialogue Challenge (Li et al., 2016, 2018) with three datasets across different difficulties for our experiments: Movie-Ticket Booking, Restaurant Reservation, Taxi Ordering. As far as the typical difficulty classification criteria (the number of slots in the user goal) (Zhao et al., 2020; Liu et al., 2021), the difficulty of each dataset is: movie = easy, rest. = middle, taxi = hard, as shown in Figure 5. The three datasets performed experimentally are human-annotated, and their statistics are shown in Table 2.

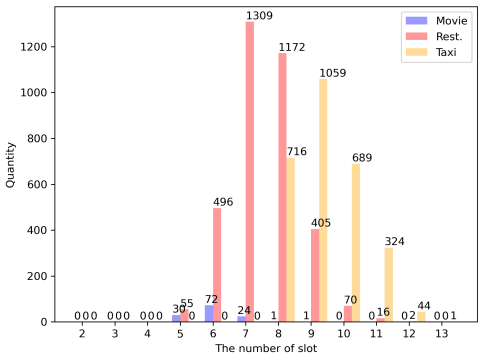


Figure 5: The distribution of the number of slots for user goals in each dataset.

Task	Intents	Slots	Dialogues	User goals	Mainly slot numbers	Domain
Movie-Ticket Booking	11	29	2890	128	5-7	1
Restaurant Reservation	11	30	4103	3525	5-12	1
Taxi Ordering	11	29	3094	2830	8-13	1

Table 2: The number of intents, slots, dialogues, user goals, and the range of slot number in user goals in three datasets.

4.2 Baselines

To verify the effectiveness and versatility of our method and its global curriculum, we conduct experiments using the following existing dialogue policies with curriculum learning and the curriculum standards it brings as benchmarks, including:

- **DQN** model is implemented using a standard DQN with only direct reinforcement learning (Mnih et al., 2015).
- **ACL-DQN(A/B/C)** model performs automatic dialogue curriculum learning by using an RL-based teacher model to change the learning order of the dialogue agents based on their learning process, which contains three schedules (Zhao et al., 2021a). Schedule A has no fixed criteria and uses only a single teacher model. Both schedule B and schedule C ensure that the teacher model develops curriculum from easy to complex, with schedule B forcing students to learn one by one and not skipping levels while Schedule C allows it.
- **HER** model performs dialogue data augmentation by segmenting successful sessions from failed dialogues and stitching them with similar successful dialogue to obtain artificial successful dialogues (Lu et al., 2019).
- **PER** model reconstructs the training process by prioritizing and replaying samples with large temporal difference (TD) errors more (Schaul et al., 2016).
- **SDPL** model uses the dialogue state differential space (ICM loss) to assess the difficulty of user goals and then scales the capacity of the training goal set proportionally as the training time increases (Liu et al., 2021).

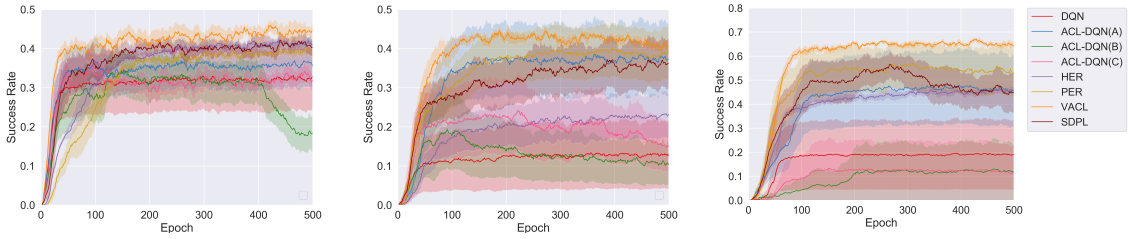
Proposed VACL

- **VACL** is the model we proposed that would apply to the varying environments, students, and support conditioning to handle changing tasks, exempt from the limitation of unable human-defined difficulty.
- **SNCL** is a variant of VACL, which replaces the Difficulty Measure with the Slot-Number based Difficulty Measure from Zhao et al. (2021a) [4].
- **VOCL** is a variant of VACL, which replaces the Training Scheduler with the sequential training, forcing students to learn from easiness to difficulty one by one.

[4] In implementing SNCL, we fix the ΔR in Eq. 2 to -1, since the slot-number based difficulty value and the ΔR do not belong to the same order of magnitude, and their difficulty directions are opposite.

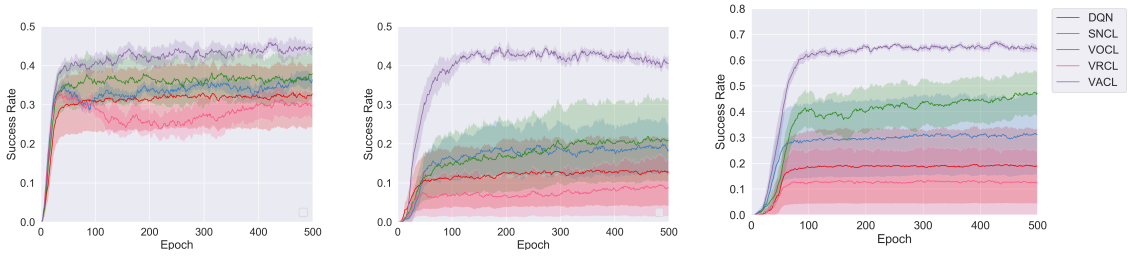
Agent	domain	Epoch = 50			Epoch = 150			Epoch = 250			Epoch = 350		
		Success	Reward	Turns	Success	Reward	Turns	Success	Reward	Turns	Success	Reward	Turns
DQN	Movie	0.2999	-19.13	32.27	0.3128	-17.61	32.29	0.3250	-15.97	31.93	0.3202	-16.67	32.18
ACL-DQN(A)		0.3453	-13.00	30.91	0.3366	-14.43	31.63	0.3665	-10.42	30.81	0.3566	-11.83	31.23
ACL-DQN(B)		0.2463	-26.46	34.04	0.3379	-14.41	31.91	0.3223	-16.58	32.51	0.3138	-17.71	32.74
ACL-DQN(C)		0.3266	-15.90	32.20	0.3167	-17.38	32.77	0.3077	-18.63	33.10	0.3184	-17.10	32.62
HER		0.1877	-34.14	35.34	0.3705	-9.69	30.30	0.3993	-6.09	30.01	0.4030	-5.65	30.01
PER		0.1360	-41.00	36.66	0.3534	-12.07	30.95	0.3693	-10.04	30.73	0.3830	-8.24	30.41
SDPL		0.3188	-16.42	31.38	0.3947	-6.60	29.94	0.4050	-5.20	29.60	0.4050	-5.28	29.76
VACL*		**0.3883**	**-7.36**	**29.94**	**0.4275**	**-2.22**	**29.05**	**0.4388**	**-0.63**	**28.57**	**0.4290**	**-2.03**	**29.02**
SNCL		0.3373	-13.87	30.72	0.3264	-15.50	31.33	0.3343	-14.73	31.70	0.3399	-14.06	31.69
VOCL		0.3511	-12.38	31.04	0.3569	-11.73	31.13	0.3642	-10.78	30.96	0.3635	-10.93	31.09
VRCL		0.3135	-17.02	31.31	0.2479	-26.24	33.98	0.2475	-26.38	34.16	0.2810	-21.88	33.20
DQN	Rest.	0.1058	-34.34	29.74	0.1103	-34.00	29.87	0.1259	-32.48	29.61	0.1315	-31.92	29.52
ACL-DQN(A)		0.1908	-25.76	27.88	0.3618	-9.01	25.14	0.3669	-8.46	24.97	0.3737	-7.80	24.85
ACL-DQN(B)		0.1605	-28.91	28.72	0.1486	-30.15	29.04	0.1169	-33.28	29.60	0.1131	-33.72	29.80
ACL-DQN(C)		0.2031	-24.72	28.02	0.2201	-23.10	27.83	0.2035	-24.87	28.36	0.2097	-24.18	28.09
HER		0.1016	-34.64	29.58	0.1943	-25.89	28.74	0.2129	-24.08	28.48	0.2109	-24.29	28.53
PER		0.2319	-21.65	27.06	0.3558	-9.58	25.21	0.3822	-6.97	24.74	0.3903	-6.18	24.63
SDPL		0.2499	-19.87	26.74	0.3171	-13.45	25.99	0.3458	-10.61	25.46	0.3432	-10.89	25.56
VACL*		**0.3002**	**-14.92**	**25.89**	**0.4326**	**-2.06**	**23.98**	**0.4378**	**-1.57**	**23.95**	**0.4256**	**-2.76**	**24.14**
SNCL		0.1291	-31.72	28.70	0.1751	-27.48	28.47	0.1840	-26.62	28.36	0.1828	-26.74	28.38
VOCL		0.1144	-33.51	29.61	0.1571	-29.39	29.05	0.1867	-26.47	28.54	0.1997	-25.16	28.27
VRCL		0.0735	-37.61	30.47	0.0718	-37.86	30.63	0.0799	-37.08	30.54	0.0739	-37.68	30.66
DQN	Taxi	0.0974	-34.96	29.47	0.1916	-25.96	28.40	0.1925	-25.85	28.35	0.1882	-26.30	28.47
ACL-DQN(A)		0.1673	-27.57	27.26	0.4375	-1.63	24.00	0.4629	0.93	23.47	0.4623	0.90	23.42
ACL-DQN(B)		0.0182	-43.10	31.49	0.0644	-38.61	30.80	0.1223	-32.93	29.89	0.1173	-33.40	29.91
ACL-DQN(C)		0.0233	-42.33	30.87	0.1312	-31.97	29.54	0.1160	-33.46	29.80	0.1207	-33.02	29.76
HER		0.2295	-22.01	27.36	0.4144	-4.12	24.83	0.4338	-2.22	24.52	0.4503	-0.56	24.17
PER		**0.3510**	**-9.77**	24.72	0.5434	8.98	21.84	0.5389	8.40	22.21	0.5510	9.68	21.82
SDPL		0.2645	-18.36	26.34	0.4903	3.49	23.27	0.5424	8.33	22.98	0.4743	1.81	23.76
VACL*		0.3396	-10.25	**23.63**	**0.6321**	**17.74**	**20.30**	**0.6600**	**20.66**	**19.49**	**0.6513**	**19.80**	**19.62**
SNCL		0.2322	-21.59	26.99	0.2908	-16.07	26.49	0.3004	-15.10	26.28	0.3006	-15.11	26.33
VOCL		0.1366	-30.34	27.27	0.3868	-6.91	25.46	0.4144	-4.14	24.86	0.4311	-2.61	24.82
VRCL		0.0784	-37.00	30.13	0.1310	-31.99	29.56	0.1227	-32.80	29.68	0.1294	-32.14	29.58

Table 1: The results of different agents at training $epoch = \{50, 150, 250, 350\}$. Each number is averaged over 10 runs, and each run is tested on 1000 dialogues. Best scores are labeled in blue. * denotes significant level $p < 0.05$ with other agents. Success: average success rate, Reward: average reward, Turn: average turn.



(a) Movie (b) Rest. (c) Taxi

Figure 3: The learning curves of different agents in Movie, Restaurant, and Taxi domains.



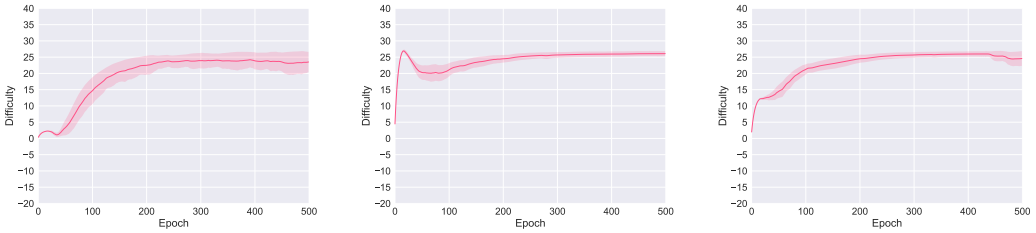
(a) Movie (b) Rest. (c) Taxi

Figure 4: The ablation experiment of two components of VACL in Movie, Restaurant, and Taxi domains.

- **VRCL**is a variant of VACL, which replaces the Training Schedule with the inversed training, forcing students to learn from difficulty to easiness one by one.

Generic Global Curriculum

- **Random Curriculum** is the user goal sam-



(a) User goal 1 (b) User goal 7 (c) User goal 13

Figure 6: The difficulty curves of three user goals in the Movie domain.

pling standard for most dialogue policy models.

- **Slot-Number Curriculum** is proposed by ACL-DQN (Zhao et al., 2021a), which takes the number of $Inform\_slot$ and $Request\_slot$ contained in user goals as the difficulty criterion [5].

- **VACL Curriculum** is a generic global curriculum obtained during the training of our proposed method.

4.3 Setup

For all models, we use a single layer perceptron with 80 neurons and *RMSprop* optimizer with a 0.001 learning rate and 16 minimum batch size. The discount factor γ is set to 0.9. The buffer size is $10k$ for all models except HER, which has a buffer size of $100k$. Whenever the current average success rate reaches the maximum and is greater than 0.3, the experience replay will be emptied. The maximum allowable number of conversations L defaults to 30, except for the movie domain which are 40. An $\epsilon - greedy$ strategy is used to achieve exploration, where $\epsilon = 0.1$. The learning rate of both the teacher model and the student model of ACL-DQN is 0.01. The slot thresholds used in ACL-DQN for dividing user goal sets are $5, 6, 7$ and $6, 7, 8$ and $8, 9, 10$, respectively. The stitch threshold in the HER follows the optimal thresholds of their papers, which is set to 0.2. **For a fair comparison, the time cost and data cost for all models are consistent**. We used the same 120 conversations to pre-filled experience replay for policy initialization and difficulty initialization and the same training epoch for policy optimization and difficulty adjustment. All results are the average values of 1000 dialogues from 10 turns with different random seeds, each run tested on 100 dialogues. The colored areas between the curves are 0.5 times the standard deviations of each episode.

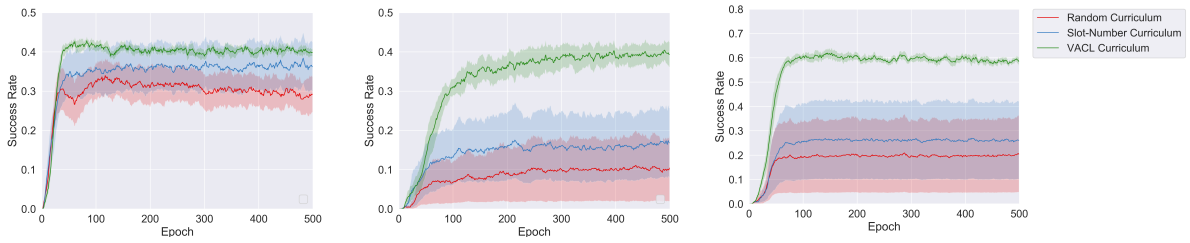
[5]It is worth noting that the difficulty of user goals does not change regardless of the case.

Index	Diffculty	Slot Number	Request_slot	Inform_slot
12	46.7863	6	\	'city': 'seattle', 'numberofpeople': '2', 'theater': 'regal meridian 16', 'starttime': '8:45 pm', 'date': 'tomorrow', 'moviename': 'the big short'
99	46.1902	5	'date': 'UNK', 'theater': 'UNK', 'starttime': 'UNK'	'numberofpeople': '4', 'moviename': 'brothers grimsby'
22	43.4684	5	'date': 'UNK', 'theater': 'UNK', 'starttime': 'UNK'	'numberofpeople': '4', 'moviename': 'zootopia'
90	29.8568	7	'date': 'UNK', 'theater': 'UNK', 'starttime': 'UNK'	'city': 'portland', 'state': 'oregon', 'numberofpeople': '4', 'moviename': 'star wars'
77	15.6063	7	'date': 'UNK', 'theater': 'UNK', 'starttime': 'UNK'	'city': 'birmingham', 'state': 'al', 'numberofpeople': '2', 'moviename': 'zootopia'
115	6.6631	6	'date': 'UNK', 'theater': 'UNK', 'starttime': 'UNK'	'state': 'california', 'numberofpeople': '4', 'moviename': 'zootopia'
92	-13.6473	7	'theater': 'UNK', 'starttime': 'UNK'	'date': 'tomorrow', 'city': 'philadelphia', 'numberofpeople': '4', 'moviename': 'deadpool', 'zip': '19101'
16	-14.8769	6	'date': 'UNK', 'theater': 'UNK', 'starttime': 'UNK'	'city': 'Petaluma', 'numberofpeople': '4', 'moviename': 'eddie the eagle'
84	-15.4901	6	\	'city': 'seattle', 'numberofpeople': '2', 'theater': 'regal meridian 16', 'starttime': '8:45 pm', 'date': 'tomorrow', 'moviename': 'hail caesar'

Table 3: Information of selected user goals in the acquired movie global curriculum.

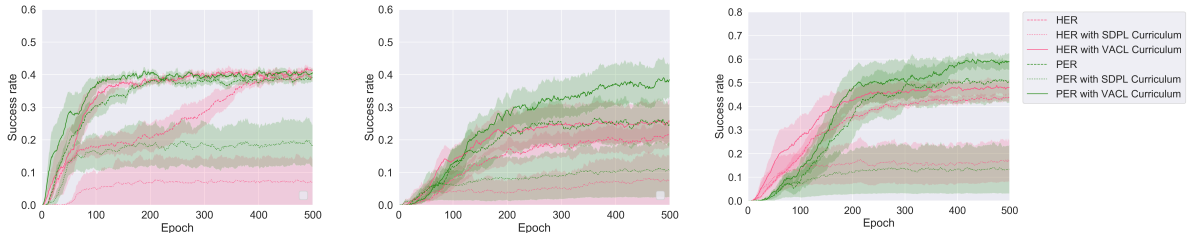
4.4 Effectiveness Evaluation

The results of our proposed VACL and the comparison models in the three domains are reported in Table 1 and Figure 3. It is obvious that the VACL brings solid improvement, affirming the effectiveness of VACL. Although PER achieves almost the second-best performance at the final epoch, it can be seen from Table 1 that PER learns slowly in the early stage. In contrast, HER has shown competitive performance in the movie domain, while it seems to have little advantage in the other two do-



(a) Movie (b) Rest. (c) Taxi

Figure 7: The learning curves of DQN agents trained with different curriculums in Movie, Restaurant, and Taxi domains.



(a) Movie (b) Rest. (c) Taxi

Figure 8: The learning curves of different agents (HER and PER) trained with our global curriculum in Movie, Restaurant, and Taxi domains.

mains. We also notice that ACL-DQN(A) performs better than ACL-DQN(B) and ACL-DQN(C). We conjecture that the difficulty metric of user goal has an extremely complex criterion, not just the slot number of user goals. Thus, ACL-DQN(A) learns more hidden difficulty rules, allowing dialogue agents to benefit more from it. It also validates our assumption about the difficulty of dialogue tasks. Although SDLP learns efficiently in the early stages, its performance is affected by the forgetting problem caused by its pacing function. Moreover, the integration of SDPL into our framework achieves better performance, which further strengthens the contribution of our approach. The experimental results are presented in Appendix .4.

4.5 Ablation Analysis

To further analyze the contribution of two components to the proposed method, we conduct the ablation test. As shown in Table 1 and Figure 4, replacing either component hugely affects the performance of dialogue policy, especially in the more difficult restaurant and taxi domains. Among them, VRCL performs worst over other models in all domains while VOCL has slight improvements over DQN and SNCL. And both of them are worse than VACL. This result implies that the validity of the training scheduler and the accuracy of our difficulty measurer. Although SNCL produces better performance than DQN, it is still limited by the fixed user goal difficulties and fails to train a good policy. The result also demonstrates that user goal difficulties evaluated by our difficulty measurer have more accurate than simply using the slot number. In conclusion, these two components benefit the VACL to a large extent.

4.6 Versatility Analysis

We suppose that our VACL can build a generic, elastic global curriculum while training a good policy, which could guide different students learning without extra re-training. To verify our conjecture, we plot and observe the curves of each user goal's difficulty during the training process. We find that almost all user goal difficulties converged, and we randomly select three user goals with 10 random seeds and draw their average difficulty curves shown in Figure 6. The results initially verify our conjecture.

The effectiveness of the global curriculum needs to be demonstrated before verifying its generality. Hence, we integrate different global curriculums into a classic representative of the DRL-based dialogue policies, DQN policy, and see if our global curriculum performs best among all of them. Figure 7 reports the effect of different global curricu-

Agent	Movie		Rest.		Taxi	
	Success	Rating	Success	Rating	Success	Rating
DQN	0.17	2.13	0.04	0.93	0.05	1.03
ACL-DQN(A)[7]	0.28	2.86	0.22	2.25	0.32	2.46
HER	0.30	2.47	0.13	1.64	0.31	1.97
PER	0.34	2.96	0.22	2.41	0.35	2.60
SDPL	0.34	2.78	0.20	2.35	0.38	2.73
VACL*	**0.39**	**3.13**	**0.28**	**2.96**	**0.44**	**3.02**
VACL-DQN	0.36	3.05	0.23	2.61	0.40	2.82

Table 4: Human evaluation of different agents in Movie, Rest. and Taxi domains. Scores with $*$ are statistically significant ($p < 0.05$).

lum on the DQN model. It is obvious that training with our VACL curriculum achieves the best performance across three domains and consistently outperforms other curriculums by a large margin. Interm of slot-number curriculum, it always improves very little.

To further glean insight regarding the generality of our global curriculum, we integrate our global curriculum and SDPL curriculum [6] with different students (HER, and PER) and see whether its performance is improved. Too many the number of training leads the performance of student model converged too quickly to clearly observe the effect of our global curriculum. *Therefore, we reduced the number of training by a factor of 5*. Figure 8 reports the learning curves of different students on our global curriculum. It is obvious that training with our global curriculum produces significant improvements while the SDPL curriculum does not.

Therefore, we can conclude that our global curriculum could be applied to guide diverse student learning without extra effort on re-training, which verifies the versatile of the VACL framework.

4.7 Global Curriculum Analysis

To analyze the factors affecting the difficulty of user goals, we further compare user goals with varying difficulties in the VACL curriculum. Taking the movie domain as an example, we randomly selected three user goals in the VACL curriculum located at the head (blue), middle-part (yellow), and tail (gray) positions respectively, and grouped their information into Table 3. It can be seen that the slot number does not fully reflect the user goal difficulty. For example, although user goal 115 increases its difficulty by adding one inform_slot over 22, user goal 77 with 7 slots is easier than user goal 84 with six slots. In addition, we also observed an interesting phenomenon that, both have the same slot-value pairs except *moviename*, user goal 12 and user goal 84 have the exact opposite difficulty, while user goal 99 and user goal 22 have approximate difficulty. We conjecture that the user goal difficulty is more relevant to the slot value in inform_slot, which determines the amount of available information. This theoretical result is further analyzed in Appendix .2.

4.8 Human Evaluation

To evaluate the feasibility of our VACL from a human perspective, we recruited real users to interact with different agents trained for 300[8] epochs using the platform and human evaluation criteria provided by MDC. The results are shown in Table 4. Consistent with simulation results, VACL also achieves better performance from a human perspective. Furthermore, we observe that the performance of VACL-DQN (DQN with our generic global curriculum) is close to VACL and outperformance of all others. It also verifies the effectiveness of our generic global curriculum. The detailed criteria for the human evaluation are shown in Appendix .5.

5 Conclusion

This paper proposes a versatile adaptive curriculum learning (VACL) framework for task-oriented dialogue policy learning. It supports evaluating the difficulty of dialogue tasks only using the learning experiences of dialogue policy and adaptively curriculum-conditioning based on the learning needs of students. This framework presents a substantial step toward applying automatic curriculum learning on dialogue policy tasks. Moreover, we explored the versatility of VACL that it succeeds in training a good dialogue policy while building a generic curriculum to guide diverse student learning without extra effort on re-training. Additionally, we analyzed the ranking of the generic global curriculum, which gave us some inspiration to refine the difficulty criteria for dialogue tasks. An important direction for our future work is how to expand our approach from single-domain tasks with different difficulties to multi-domain tasks, e.g. MultiWoz (Budzianowski et al., 2018). This paper is an essential cornerstone for such an investigation.

[6] It is worth noting that SDPL inability to produce a stable, generic curriculum. To test the point, we choose the SDPL model trained after 500 epochs in three domains to evaluate the difficulty of each user goal and rank them to form an SDPL curriculum.

[8] From the above experiments, it is observed that all models have converged and achieved optimal performance at 300 epoch.

6 Acknowledgments

We would like to thank the reviewers for their comments and efforts towards improving our paper. And we would like to acknowledge volunteers of the South China University of Technology who help us with the human experiments. This work was supported by the Key-Area Research and Development Program of Guangdong Province, China (Grant No.2019B0101540042) and the Natural Science Foundation of Guangdong Province, China (Grant No.2019A1515011792).

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Yang Fan, Fei Tian, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2018. Learning to teach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Milica Gasic, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve J. Young. 2013. Online policy optimisation of bayesian spoken dialogue systems via human interaction. In *ICASSP*, pages 8367–8371. IEEE.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1311–1320. PMLR.

Michael Cerny Green, Benjamin Sergent, Pushyami Shandilya, and Vibhor Kumar. 2019. Evolutionarily-curated curriculum learning for deep reinforcement learning agents. *CoRR*, abs/1901.05431.

John Hammersley. 2013. *Monte carlo methods*. Springer Science & Business Media.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. 2015. Self-paced curriculum learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2694–2700. AAAI Press.

Tae-Hoon Kim and Jonghyun Choi. 2018. Screenernet: Learning curriculum for neural networks. *CoRR*, abs/1801.00904.

Xiujun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. 2016. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688*.

Xiujun Li, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

Sihong Liu, Jinchao Zhang, Keqing He, Weiran Xu, and Jie Zhou. 2021. Scheduled dialog policy learning: An automatic curriculum learning framework for task-oriented dialog system. In *Findings of ACL-IJCNLP 2021*, pages 1091–1102.

Keting Lu, Shiqi Zhang, and Xiaoping Chen. 2019. Goal-oriented dialogue policy learning from failures. In *AAAI*, pages 2596–2603. AAAI Press.

Nurul Lubis, Christian Geishauser, Michael Heck, Hsien-Chin Lin, Marco Moresi, Carel van Niekerk, and Milica Gasic. 2020. LAVA: latent action spaces via variational auto-encoding for dialogue policy optimization. In *COLING*, pages 465–479. International Committee on Computational Linguistics.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *J. Mach. Learn. Res.*, 21:181:1–181:50.

Sanmit Narvekar and Peter Stone. 2019. Learning curriculum policies for reinforcement learning. In *AAMAS*, pages 25–33. International Foundation for Autonomous Agents and Multiagent Systems.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. 2020. Automatic curriculum learning for deep RL: A short survey. In *IJCAI*, pages 4819–4825. ijcai.org.

Sébastien Racanière, Andrew K. Lampinen, Adam Santoro, David P. Reichert, Vlad Firoiu, and Timothy P. Lillicrap. 2019. Automated curricula through setter-solver interactions. *CoRR*, abs/1909.12892.

Zhipeng Ren, Daoyi Dong, Huaxiong Li, and Chunlin Chen. 2018. Self-paced prioritized curriculum learning with coverage penalty in deep reinforcement learning. *IEEE Trans. Neural Networks Learn. Syst.*, 29(6):2216–2226.

Jost Schatzmann and Steve J. Young. 2009. The hidden agenda user simulation model. *IEEE Trans. Speech Audio Process.*, 17(4):733–747.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized experience replay. In *ICLR*.

Oliver G. Selfridge, Richard S. Sutton, and Andrew G. Barto. 1985. Training and tracking in robotics. In *IJCAI*, pages 670–672. Morgan Kaufmann.

Pei-Hao Su, Milica Gasic, and Steve J. Young. 2018. Reward estimation for dialogue policy optimisation. *Comput. Speech Lang.*, 51:24–43.

Maxwell Svetlik, Matteo Leonetti, Jivko Sinapov, Rishi Shah, Nick Walker, and Peter Stone. 2017. Automatic curriculum graph generation for reinforcement learning agents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2590–2596. AAAI Press.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2021. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. In *ICLR*. OpenReview.net.

Daphna Weinshall and Dan Amir. 2018. Theory of curriculum learning, with convex loss functions. *CoRR*, abs/1812.03472.

Daphna Weinshall, Gad Cohen, and Dan Amir. 2018. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5235–5243. PMLR.

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

Zhirui Zhang, Xiujun Li, Jianfeng Gao, and Enhong Chen. 2019. Budgeted policy learning for task-oriented dialogue systems. In *ACL*, pages 3742–3751. Association for Computational Linguistics.

Yangyang Zhao, Zhenyu Wang, and Zhenhua Huang. 2021a. Automatic curriculum learning with over-repetition penalty for dialogue policy learning. In *AAAI*, pages 14540–14548. AAAI Press.

Yangyang Zhao, Zhenyu Wang, Kai Yin, Rui Zhang, Zhenhua Huang, and Pei Wang. 2020. Dynamic reward-based dueling deep dyna-q: Robust policy learning in noisy environments. In *AAAI*, pages 9676–9684. AAAI Press.

Yangyang Zhao, Zhenyu Wang, Changxi Zhu, and Shihan Wang. 2021b. Efficient dialogue complementary policy learning via deep q-network policy and episodic memory policy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4311–4323.

Rounds	Speaker	Content	Number of matching entries
\	\	\	991
1	U	Can I get tickets for the big short tomorrow?	4
2	S	Which city would you like?	4
3	U	I want to watch at seattle.	4
4	S	What time would you like to see it?	4
5	U	I want to watch at 8:45 pm.	4
6	S	regal meridian 16 is available.	4
7	U	I want 2 tickets please!	4
8	S	Great - I was able to purchase 2 tickets for you to see the big short tomorrow at regal merdian 16 theater in seattle at 8:45 pm.	4
9	U	Thank you.	4
10	S	Thank you.	4
Success			

Table 5: A typical example of a conversation for user goal 12.

Rounds	Speaker	Content	Number of matching entries
\	\	\	991
1	U	Can I get some tickets for hail caesar tomorrow?	10
2	S	Which city would you like?	10
3	U	I want to watch at seattle.	10
4	S	hail caesar is available.	10
5	U	I want 2 tickets please!	7
6	S	regal meridian 16 is available.	7
7	U	Could you help me to book the tickets?	7
8	S	Great, I was able to purchase 2 tickets for you to watch hail caesar tomorrow at regal meridian 16 in seattle at 10:00 pm.	7
9	U	Oh, sorry! This is the wrong ticket!	7
10	S	Great, I was able to purchase 2 tickest for you to watch hail caesar tomorrow at regal meridian 16 in seattle at 10:00 pm.	7
11	U	Oh, sorry! This is the wrong ticket!	7
....		....	
39	S	Great, I was able to purchase 2 tickest for you to watch hail caesar tomorrow at regal meridian 16 in seattle at 10:00 pm.	7
40	U	Oh, sorry! This is the wrong ticket!	7
Failure			

Table 6: A typical example of a conversation for user goal 84.

.1 User goals

The whole dialogue behavior of the user simulator is guided by a user goal, which ensures rationality, coherence, and consistency of the dialogue. During each conversation, the user simulator holds a fixed user goal to interact with the student model. This user goal describes the needs of users and explains the purpose of this dialogue. Generally, a user goal G includes a set of constraints C (also refer to $Inform\_slot$) and a set of requests R (also refer to $Request\_slot$), where C denotes the information constrained by the user and R denotes the information required by the user (Schatzmann and Young, 2009).

Taking a train-ticket booking as an example, the user's goal is to inquire about the departure time and arrival time of today's trains from Guangzhou to Shenzhen, where the user goal G is in the following form:

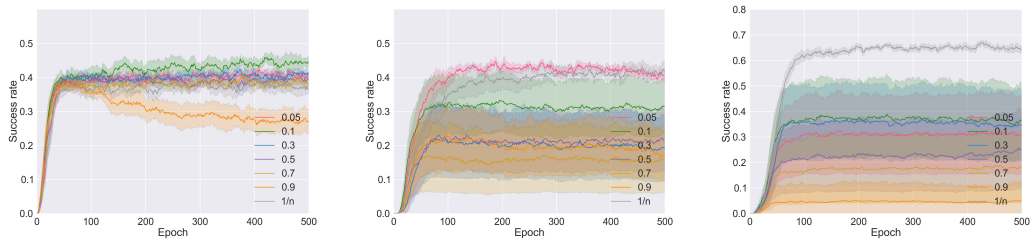
$$\textbf{Goal} = \left(C = \begin{bmatrix} location\_from = Guangzhou \\ location\_to = Shenzhen \\ date = Today \end{bmatrix}, R = \begin{bmatrix} departure\_time = ? \\ arrival\_time = ? \end{bmatrix} \right) \tag{4}$$

The condition that dialogue is considered to be successful is if and only if all the information provided by the user is identified accurately, the information provided by the agent meets all the constraints of the user, and a train-ticket that satisfies the above conditions is successfully reserved.

.2 Explanation of the ranking of the global curriculum

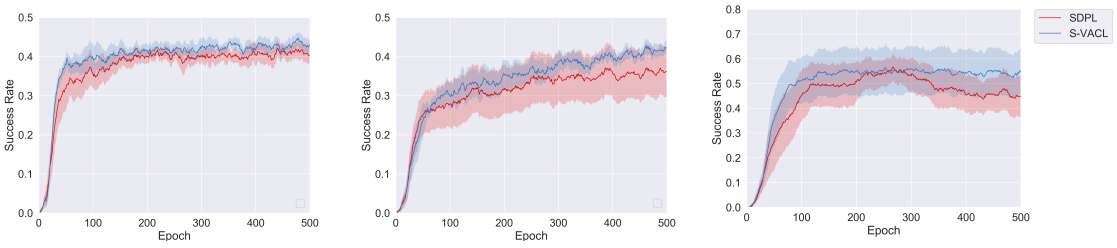
Different slots in the user goal carry different amounts of information, affecting the difficulty of user goals. User goals with more information slots (e.g., user goal 12) help the dialogue agents to locate entries in the knowledge base that satisfy the user's needs faster, thus they are. In contrast, user goals with less information slots (e.g., user goal 84) require more turns to gradually exclude unmatched entries. Therefore, such user goals are harder because dialogue agents get more penalties and even fail because of reaching the maximum number of rounds.

Take a typical conversation with user goals 84 and 12 as an example, where *U* represents the user and *S* represents the dialogue agent. As shown in Table 5 and 6, in addition to the conversation content, we also record the change in the number of movie entries that matched the current constraints in the movie knowledge base. It reflects that the amount of information of each utterance, e.g., *the big short* (movie name) and *tomorrow* (date) in the first utterance of user goal 12 directly helps the dialogue agent to filter out 4 entries from 991 entries that match all the user constraint, and after subsequent checking, the conversation is successful. Although the different slot values *hail caesar* (movie name) and *tomorrow* (date) also help dialogue agents to quickly narrow down from 991 entries, it needs more information to continue to exclude unmatched entries. In most cases, the conversation will fail due to insufficient information and reach the maximum number of rounds. Of course, there are few successful conversations, but most of them are failed examples because it reached the maximum number of rounds. Therefore, we selected typical failed examples to demonstrate. It explains that the user goals (e.g., user goal 12) with



(a) Movie (b) Rest. (c) Taxi

Figure 9: The effect of different α values on performance in Movie, Restaurant, and Taxi domains.



(a) Movie (b) Rest. (c) Taxi

Figure 10: The learning curves of S-VACL (integrating SDPL into VACL framework) agents in Movie, Restaurant, and Taxi domains.

more informative slots are easier.

.3 Training with varying values of α

α in Eq.2 controls the updating rate of the user goal difficulty. Intuitively, an $alpha$ that is too small makes it difficult to assess accurate curriculum difficulties quickly, while an $alpha$ that is too large is prone to overkill and makes curriculum difficulties amplitude too large to converge. Thus, we examine the effect of varying $alpha$ on VACL performance, which has an important reference for VACL practitioners. Figure 9 reports the experimental result in three domains. It is evident in three domains that both too large and too small α hurt the learning of dialogue policies, which is consistent with our expectations. It is worth noting that in all experiments, the α for the movie, restaurant, and taxi domains defaults to 0.1, 0.05, and $\frac{1}{n}$ respectively unless otherwise stated.

.4 Integration of SDPL into VACL framework

The pacing function in SDPL simply scales the capacity of the training goal set proportionally as the training time increases. Such a way ignores the learning need of dialogue policies, resulting in forgotten problems and performance impairment. In contrast, our training scheduler takes this important feedback into account, allowing free scheduling and achieving better results. To verify this conjecture, we integrate SDPL into our framework, namely S-VACL, and accordingly the R_i and ΔR in the training scheduler of VACL are modified to correspond to its difficulty evaluation criterion. The results are shown in Fig 10. It can be seen that the training scheduler in VACL framework is more flexible by considering the learning needs of dialogue policies to achieve better results. Therefore, we can conclude that our VACL approach achieves an optimal combination of difficulty measurer and training scheduler that is not only applicable to other reinforcement learning algorithms but also adaptable to other course learning methods.

.5 Detailed criteria for the human evaluation

For human evaluation, we recruited 36 volunteers through our labs. In each conversation, users randomly select a user goal and interact with 6 anonymous systems respectively. Each user needs to interact effectively with each system 50 times. At the end of conversations, users are required to provide feedback on whether the conversation was successful and to rate the quality of conversation on a score of 1 to 5. The quality of the conversation is affected by its degree of task completion, coherence, and naturalness. The higher the score, the better the quality of conversation.