Cold-started Curriculum Learning for Task-oriented Dialogue Policy

1st Hui Zhu
College of economics and trade
Guangdong Mechanical & Electrical Polytechnic
Guangzhou, China
1556164350@qq.com

2nd Yangyang Zhao*
School of Software Engineering
South China University of Technology
Guangzhou, China
msyyz@mail.scut.edu.cn

3rd Hua Qin*
School of Software Engineering
South China University of Technology
Guangzhou, China
qhscut@163.com

2021 IEEE International Conference on e-Business Engineering (ICEBE) | 978-1-6654-4418-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICEBE52470.2021.00015

***Abstract*—Training a satisfactory dialogue policy via Reinforcement Learning (RL) requires significant interaction costs because of delayed and sparse rewards in task-oriented dialogue tasks. Researching how to conduct efficient dialogue policy learning in environments where rewards are delayed and sparse is essential. Existing approaches obtain more positive rewards by incorporating an RL-based teacher model to customize a curriculum that matches the ability of dialogue policies for curriculum learning. However, such a teacher model still retains the disadvantage of reinforcement learning has expensive training costs. Therefore, we develop a novel framework, cold-start curriculum learning (CCL), for task-oriented dialogue policy learning that does not require any training cost for curriculum schedule. Besides, it adaptively adjusts the difficulty of the user goals and selects the next goal based on the feedback of students. Experiments show that the CCL significantly improves the effectiveness of dialogue tasks without any training cost for curriculum schedule.**

***Index Terms*—Dialogue Policy; Reinforcement Learning; Curriculum Learning; Cold-started**

I. Introduction

Learning dialogue policies are typically formulated as a reinforcement learning (RL) problem [1]–[7]. Tasks with dense rewards can help RL build effective strategies quickly [8]. However, RL-based policy learning is usually performed through delayed and sparse rewards from environments (users or user simulators) [9]. It leads to inefficient dialogue policy learning that requires a large number of conversations with environments to converge. Therefore, it is essential to study how to perform efficient dialogue learning in a reward-sparse environment.

[10] argued that the cause of the problem is mainly due to the mismatch between the selected user goals and the dialog agent capabilities. They proposed the Automatic Curriculum Learning-based Deep Q-Network (ACL-DQN) method, in which student agents (dialogue policies) learning in a tailors curriculum provided by an RL-based teacher. Experiments demonstrate learning with a suitable curriculum the training efficiency of dialogue policies and make the dialogue policies obtain positive rewards quickly. However, teachers who rely on reinforcement learning has inevitable training cost for curriculum schedule and their equipped curriculum schedules are fixed that does not adjust as the abilities of students.

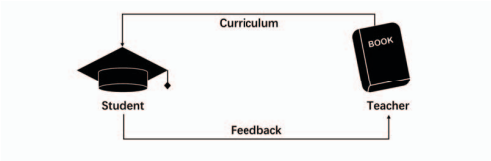


Fig. 1. The interactive process between teachers and students.

We explored cold-start curriculum learning (CCL) for task-oriented dialogue policy learning that does not require any training cost for curriculum schedule. As illustrate in Figure 1, CCL adaptively adjusts the difficulty of the user goals and selects the next goal based on the feedback of students. Specifically, the teacher model adjusts the user goal difficulty based on the results of the selected user goal. User goals that are successfully accomplished are relatively easier, while failed user goals are relatively more difficult. Similarly, the current level of the student can be expressed according to its performance in the recent period. The more successful user goals, the greater the student progress, and the harder the next goal should be.The teacher model rearrange the curriculum based on the situation of the students in completing the dialogue in the most recent period of time. Sufficient experiments have demonstrated that the CCL significantly improves the efficiency of dialogue policy learning and achieves better performance without any training cost for curriculum schedule. In summary, our contributions are as follows:

- We propose cold-start curriculum learning (CCL), which does not require any training cost for curriculum sched-

This work was supported by the 13th five year plan project of philosophy and social sciences of Guangdong Province, "Research on Reverse Technology Spillover, independent innovation and China's economic growth effect under the condition of open economy" (No.GD20XYJ20); key soft science project of Guangdong electromechanical vocational and Technical College: analysis of Sino US import and export commodity structure under the background of trade war upgrading (No.YJZD20210002).

*Corresponding author
*Corresponding author

978-1-6654-4418-7/21/$31.00 ©2021 IEEE
DOI 10.1109/ICEBE52470.2021.00015

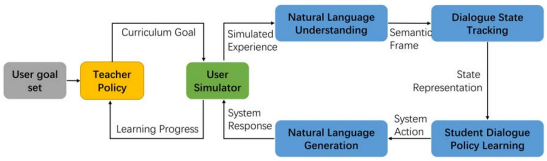


Fig. 2. Ilustration of the proposed CCL dialogue system framework.

ule, and adaptively adjusts the difficulty of the user goals based on the feedback of students.

- We validate the superior performance of CCL by building dialogue agents for the three public dialogue tasks. The efficiency of CCL is verified by simulation and human evaluations.

II. Method

As illustrate in Figure 2, the CCL dialogue system framework consists of six modules: (1) an LSTM-based natural language understanding (NLU) module [11] for identifying user intents and extracting associated slots; (2) a dialogue state tracker for tracking dialogue states [12]; (3) a student dialogue policy that selects next action based on the current state (the student model in our CCL method); (4) a natural language generation (NLG) module for generating natural language response based on current action [13]; (5) a user simulator for replacing the real user to interact with dialogue system around the user goal selected by the teacher model, where the dialogue system knows nothing about this user goal while the user simulator knows [14]; and (6)a teacher model arrange suitable curriculum for dialogue policy learning according to the feedback of the student model.

As shown in Figure 1, we proposed *Cold-started Curriculum Learning* (CCL), an adaptive curriculum learning framework, where the teacher adaptively arranges curriculums according to students' feedback to achieve student learning from easy to complex without any training cost for curriculum schedule. In this section, we introduce two components of CCL: i) *Teacher model* defines the difficulty criterion of students' goals and selects the appropriate goals for students according to students' feedback. ii) *Student model* leans from easy to complex in line with the curriculum arranged by teachers and give feedback to the teacher for adjusting.

A. Teacher Model

1) Difficulty Evaluation: 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 and a set of requests R, where C denotes the information constrained by the user and R denotes the information required by the user [8], [15].

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:

$$\textbf{Goal} = \left(C = \begin{bmatrix} location\_from = Guangzhou \\ location\_to = Shenzhen \\ date = Today \end{bmatrix}, \right. \\ \left. R = \begin{bmatrix} departure\_time = ? \\ arrival\_time = ? \end{bmatrix} \right) \tag{1}$$

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. Hence, the fewer constraints and requests contained in C and R, the fewer actions the agent needs to perform to complete G so that it is less prone to make mistakes. Conversely, the more information contained, the more actions it needs to perform, and it is more prone to make mistakes, leading to dialogue failure. Therefore, the length of C and R can measure the difficulty of the user goal:

Definition 1: ***Goal Difficulty*** $D(G_i) = |C_i| + |R_i|$, where $|C_i|$ is the number of inform slots in user goal G_i, and $|R_i|$ is the number of request slots in user goal G_i.

Consider the user goal G in Equ.1, C contains three inform slots and R contains two request slots. Generally, in order to achieve the goal, the system needs to obtain the values of three inform slots and provide the user with the information of two request slots. Therefore, the difficulty of this user goal G is 5, $D(G) = |C| + |R| = 3 + 2 = 5$.

Of course, the difficulty of user goals should not be set in stone. Their difficulty should be adjusted based on the completion degree of students for user goals. Therefore, the difficulty adjustment for user goals is defined as follows:

Definition 2: ***Difficulty Adjustment*** Intuitively, user goals successfully completed are relatively easy, while failed user goals are more difficult. When a student successfully completes a user goal, the difficulty of this user goal should be reduced, $D(G) = D(G) - 1$. Conversely, when a student fails to complete a user goal, the difficulty of this user goal should be increased, $D(G) = D(G) + 1$.

As in Equ.1, the current difficulty of this user goal is 5. When the student has successfully completed this user goal, its difficulty is changed to 4. Conversely, when the user goal fails, its difficulty is changed to 6.

2) Curriculum Arrangement: In human education, students learn from the easiest and then gradually increase the difficulty of the course. Once the user goals are ranked according to their difficulty, the teacher model selects the easiest user goal

for students to start training, and then gradually increases the difficulty of user goals. Considering that students may progress more, too simple user goals do not match the students' progress. Also, the current level of the student can be expressed according to its performance in the recent period. Therefore, the teacher model needs to rearrange the curriculum based on the situation of the students in completing the dialogue in the most recent period of time.

We choose the success or failure of the dialogue of nearly ten rounds of student models as the basis for the next user goal selection. The closer the dialogue results to the current round, the greater the impact, and the weight of the last n rounds, $w_n = (\frac{1}{2})^{(n-1)}$. The teacher schedules the next user goal based on the following:

$$I(G') = I(G) + \frac{\sum_{n=1}^{10} w_n \cdot d_n}{\sum_{n=1}^{10} w_n} \cdot L \tag{2}$$

where I(G) is the index of the current user goal and L is the maximum number of dialogue rounds, d_n is -1 if the last n rounds dialog is failure, otherwise d_n is +1 if the last n rounds dialog is successful.

B. Student Model

The goal of students is to achieve a specific user goal through a sequence of actions with a user simulator, which can be considered as a Markov decision process (MDP), which is well-suitable solved using reinforcement learning. This paper utilizes a deep Q-network (DQN) [16] to train the student model based on experiences stored in the student experience replay buffer D :

- The state s_t consists of five components: 1) one-hot representations of the current user action and mentioned slots; 2) one-hot representations of last system action and mentioned slots; 3) the belief distribution of possible value for each slot; 4) both a scalar and one-hot representation of current turn number; and 5) a scalar representation indicating the number of results which can be found in the database according to current search constraints.
- The action a_t corresponds pre-defined action set, such as request, inform, confirm_question, confirm_answer, etc.
- The reward r: once a dialogue reaches the successful, the student agent receives a big bonus $2L$. Otherwise, it receives $-L$. In each turn, the student agent receives a fixed reward -1 to encourage shorter dialogues.

At each step, according to the observed state s, the student model selects a random action with the probability ϵ or otherwise the greedy policy a with maximizing $Q(s, a; \theta)$.The student agent then receives reward r,and updates the state to s'. Finally, we store the experience tuple (s, a, r, s') in the student experience replay buffer D^S. This cycle continues until the dialogue terminates.

We improve the value function $Q(s, a, \theta)$ by adjusting θ to minimize the mean-squared loss function as follows:

$$\begin{aligned} \mathcal{L}(\theta^S) &= \mathbb{E}_{(s,a,r,s')\sim D^S}[(y_i - Q(s, a; \theta))^2] \\ y_i &= r + \gamma \max_{a'} Q'(s', a'; \theta^{'}) \end{aligned} \tag{3}$$

where $\gamma \in [0, 1]$ is a discount factor, and $Q'(\cdot)$ is the target value function that is only updated periodically. $Q(\cdot)$ can be optimized through $\nabla_{\theta^S} \mathcal{L}(\theta^S)$ by back-propagation and mini-batch gradient descent.

III. Experiments

Experiments have been conducted to evaluate the efficiency of CCL-based dialogue policies, in two settings: simulation and human evaluation.

A. Dataset

Our CCL was evaluated on three tasks in both simulation and human-in-the-loop settings: movie-ticket booking, restaurant reservation, and taxi ordering. Raw conversational data in three tasks was provided by Microsoft Dialogue Challenge [14], [17][1] with well annotations. The annotated datasets are shown in Table II.

Task	Intents	Slots	Dialogues
Movie-Ticket Booking	11	29	2890
Restaurant Reservation	11	30	4103
Taxi Ordering	11	29	3094

TABLE II
THE NUMBER OF INTENTS, SLOTS, DIALOGUES AND USER GOALS IN THREE DATASETS.

B. Baselines

To verify the efficiency of CCL, we developed different version of task-oriented dialogue agents as baselines to compare, including:

- The **DQN** agent takes the user goal randomly sampled by the user simulator for learning [16].
- The proposed **CCL w/o DA** agent takes the curriculum specified by the teacher model for curriculum policy learning, and the difficulty of the user goals does not change with its ability.
- The proposed **CCL w/o CA** agent takes the curriculum provided by the teacher model for sequence policy learning without curriculum arrangement.
- The proposed **CCL** agent takes the curriculum specified by the teacher model for curriculum policy learning.

C. Setup

For all the models, we use MLPs to parameterize the value networks $Q(\cdot)$ with one hidden layer of size 80 and $RMSprop$ optimizer [18]. ϵ-greedy is always applied for exploration with $\epsilon = 0.1$. We set the discount factor $\gamma = 0.9$. The buffer size of D is set to $10k$. The batch size is 16, and the learning rate is 0.001. We applied gradient clipping on all the model parameters with a maximum norm of 1 to

[1]https://github.com/xiul-msr/e2e_dialog_challenge

Agent	Task	Epoch = 100			Epoch = 300			Epoch = 500		
		Success	Reward	Turns	Success	Reward	Turns	Success	Reward	Turns
DQN	**Movie**	0.308	32.31	-18.15	0.328	31.83	-15.56	0.328	31.92	-15.64
CCL w/o CA		0.344	31.13	-13.23	0.353	31.25	-12.31	0.350	31.36	-12.68
CCL w/o DA		0.331	31.63	-15.05	0.332	31.87	-15.04	0.320	32.38	-16.82
CCL		0.395	29.63	-6.362	0.407	29.54	-4.925	0.415	29.48	-3.925
DQN	**Rest.**	0.091	30.26	-35.91	0.105	30.07	-34.63	0.106	30.04	-34.47
CCL w/o CA		0.138	29.22	-31.20	0.151	29.16	-30.01	0.153	29.12	-29.76
CCL w/o DA		0.129	29.49	-32.18	0.158	28.70	-29.15	0.156	28.87	-29.43
CCL		0.166	28.82	-28.50	0.217	27.85	-23.43	0.217	27.91	-23.44
DQN	**Taxi.**	0.185	28.49	-26.57	0.186	28.48	-26.48	0.189	28.45	-26.18
CCL w/o CA		0.246	27.28	-20.48	0.255	27.16	-19.64	0.267	27.03	-18.47
CCL w/o DA		0.243	26.71	-20.46	0.242	26.66	-20.49	0.247	25.96	-19.79
CCL		0.318	26.07	-13.41	0.323	25.80	-12.82	0.329	25.80	-12.28

TABLE I

RESULT OF DIFFERENT AGENTS AT $epoch = \{100, 200, 300\}$. EACH NUMBER IS AVERAGED OVER 10 TURNS, EACH RUN TESTED ON 100 DIALOGUES. SUCCESS: SUCCESS RATE, REWARD: AVERAGE REWARD, TURNS: AVERAGE TURNS. EVALUATED AT THE SAME EPOCH, CCL OUTPERFORMS OTHER BASELINES. THE BEST SCORES ARE LABELED IN BOLD.

prevent gradient explosion. The target network is updated at the beginning of each training episode. The maximum length of a simulated dialogue L is 40 turns. The dialogues are counted as failed, if exceeding the maximum length of turns. For training the agents more efficiently, we utilized a variant of imitation learning, called Reply Buffer Spiking (RBS) [1] at the beginning stage to build a naive but occasionally successful rule-based agent based on the human conversational dataset. We also pre-filled the real experience replay buffer B^u with 100 dialogues before training for all the variants of agents.

D. Simulation Evaluation

Figure 3 depicts the learning curve of all models under the three domains. The results show that the proposed CCL method significantly outperformed all other baselines. The validity and feasibility of our method are hereby justly verified. In addition, we also observed that $CCLw/oCA$ and $CCLw/oDA$ are also better than baseline DQN. In this case, $CCLw/oCA$ compared to DQN is trained in the sequence using the curriculum provided by the teacher model, which demonstrates the effectiveness of our difficulty evaluation module. $CCLw/oDA$ compared to DQN is trained with a teacher model for curriculum arrangement based on student completion on fixed user goal difficulty, which illustrates the effectiveness of our curriculum arrangement module. Further, the $CCLw/oCA$ and $CCLw/oDA$ performed worse than the CCL model, validating our conclusions above.

As described in the previous section, our proposed model is mainly used to facilitate agents to obtain positive rewards faster to improve their learning efficiency. Table I shows the performance details including success rate (Figure 3), average rewards (Figure 4), and the average number of rounds (Figure 5) for each model at 100epoch,300 epoch,500 epoch, respectively. As can be seen from table I, our model facilitates the acquisition of positive rewards by agents while improving the efficiency of the agents.

Agent	Movie		Rest.		Taxi	
	Success	Rating	Success	Rating	Success	Rating
DQN	0.27	1.98	0.04	1.23	0.15	1.74
CCL w/o DA	0.29	2.07	0.16	1.43	0.22	1.89
CCL w/o CA	0.30	2.32	0.17	2.02	0.28	2.00
CCL*	**0.46**	**2.87**	**0.31**	**2.62**	**0.39**	**2.37**

TABLE III

HUMAN EVALUATION OF DIFFERENT AGENTS IN MOVIE, REST. AND TAXI DOMAINS.

E. Human Evaluation

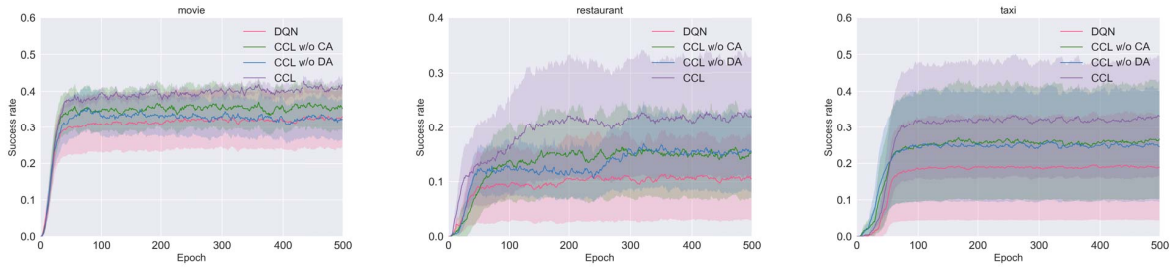
We recruited real users to evaluate different systems by interacting with different systems, without knowing which the agent is hidden from the users. All evaluated agents have been trained for 300 epochs. At the beginning of each dialogue session,the user randomly picked one of the agents to converse using a random user goal. The user can terminate the dialogue at any time if the user deems that the dialogue is too procrastinated and it is almost impossible to achieve their goal. At the end of the conversation, users are required to provide feedback on whether the conversation is successful and the a score of 1 to 5 [2] for the naturalness, coherence, and task completion ability of the agent .

As illustrated in Table III, the CCL is significantly outperforms other agents, which is consistent with what we have observed in simulation evaluation.

IV. CONCLUSION

In this paper, we proposed a novel model, Cold-started Curriculum Learning, which adaptively adjusts the difficulty of the user goals and arranges a suitable curriculum based on the feedback of students without any training cost for curriculum design. We designed a teacher model to tailor a curriculum for student training, achieving positive rewards faster. We validated the effectiveness of the proposed CCL on three publicly available dialogue datasets. In the future,

[2] 5 is the best, 1 is the worst

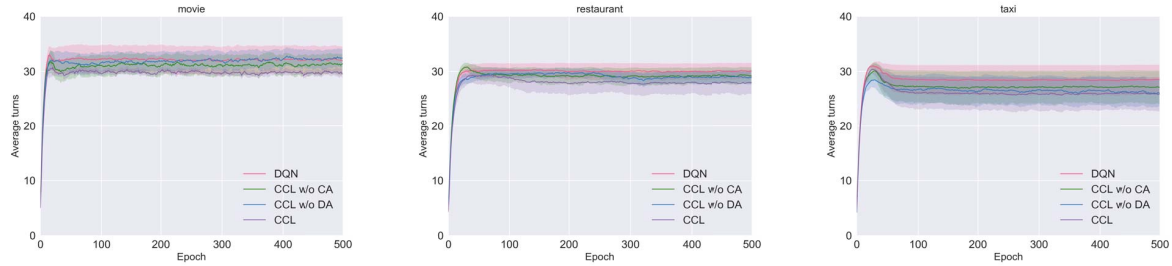


(a) Movie (b) Rest. (c) Taxi.

Fig. 3. The learning curves of all agents in movie, restaurant, and taxi domains.

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

Fig. 4. The average_rewards curves of all agents in movie, restaurant, and taxi domains.



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

Fig. 5. The average_turns curves of all agents in movie, restaurant, and taxi domains.

we plan to further evaluate the validity of our approach by equipping it as a dynamic curriculum to different curriculum learning methods.

ACKNOWLEDGMENT

We are grateful to the students who helped us with human evaluation. And we also would like to thanks the reviewers for their comments and efforts towards improving our manuscript.

REFERENCES

[1] Z. C. Lipton, J. Gao, L. Li, X. Li, F. Ahmed, and L. Deng, "Efficient exploration for dialog policy learning with deep BBQ networks \& replay buffer spiking," *CoRR*, vol. abs/1608.05081, 2016. [Online]. Available: http://arxiv.org/abs/1608.05081

[2] P. Su, M. Gasic, N. Mrksic, L. M. Rojas-Barahona, S. Ultes, D. Vandyke, T. Wen, and S. J. Young, "On-line active reward learning for policy optimisation in spoken dialogue systems," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

[3] S. Su, X. Li, J. Gao, J. Liu, and Y. Chen, "Discriminative deep dyna-q: Robust planning for dialogue policy learning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2018, pp. 3813–3823.

[4] B. Peng, X. Li, J. Gao, J. Liu, and K. Wong, "Deep dyna-q: Integrating planning for task-completion dialogue policy learning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, I. Gurevych and Y. Miyao, Eds. Association for Computational Linguistics, 2018, pp. 2182–2192.

[5] Y. Zhao, Z. Wang, P. Wang, T. Yang, R. Zhang, and K. Yin, "A survey on task-oriented dialogue systems," *Chinses journal of computers*, vol. 43,

no. 10, pp. 1862–1896, 2020.

[6] Y. Wu, X. Li, J. Liu, J. Gao, and Y. Yang, "Switch-based active deep dyna-q: Efficient adaptive planning for task-completion dialogue policy learning," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 7289–7296.

[7] Y. Zhao, Z. Wang, C. Zhu, and S. Wang, "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, EMNLP 2021*. Association for Computational Linguistics, 2021.

[8] Y. Zhao, Z. Wang, K. Yin, R. Zhang, Z. Huang, and P. Wang, "Dynamic reward-based dueling deep dyna-q: Robust policy learning in noisy environments," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 9676–9684. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6516

[9] K. Lu, S. Zhang, and X. Chen, "Goal-oriented dialogue policy learning from failures," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2596–2603.

[10] Y. Zhao, Z. Wang, and Z. Huang, "Automatic curriculum learning with over-repetition penalty for dialogue policy learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 540–14 548.

[11] D. Hakkani-Tür, G. Tür, A. Celikyilmaz, Y. Chen, J. Gao, L. Deng, and Y. Wang, "Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 715–719. [Online]. Available: https://doi.org/10.21437/Interspeech.2016-402

[12] N. Mrksic, D. Ó. Séaghdha, T. Wen, B. Thomson, and S. J. Young, "Neural belief tracker: Data-driven dialogue state tracking," *CoRR*, vol. abs/1606.03777, 2016. [Online]. Available: http://arxiv.org/abs/1606.03777

[13] T. Wen, M. Gasic, N. Mrksic, P. Su, D. Vandyke, and S. J. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, Eds. The Association for Computational Linguistics, 2015, pp. 1711–1721. [Online]. Available: https://doi.org/10.18653/v1/d15-1199

[14] X. Li, Z. C. Lipton, B. Dhingra, L. Li, J. Gao, and Y.-N. Chen, "A user simulator for task-completion dialogues," *arXiv preprint arXiv:1612.05688*, 2016.

[15] J. Schatzmann and S. J. Young, "The hidden agenda user simulation model," *IEEE Trans. Speech Audio Process.*, vol. 17, no. 4, pp. 733–747, 2009. [Online]. Available: https://doi.org/10.1109/TASL.2008.2012071

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nat.*, vol. 518, no. 7540, pp. 529–533, 2015. [Online]. Available: https://doi.org/10.1038/nature14236

[17] X. Li, S. Panda, J. Liu, and J. Gao, "Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems," *arXiv preprint arXiv:1807.11125*, 2018.

[18] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," *Cited on*, vol. 14, no. 8, p. 2, 2012.