Bootstrapped Policy Learning: Goal Shaping for Efficient Task-oriented Dialogue Policy Learning

Extended Abstract

Yangyang Zhao
Changsha University of Science and Technology, Utrecht University
Changsha, China
msyyz@mail.scut.edu.cn

Mehdi Dastani
Utrecht University
Utrecht, Netherlands
M.M.Dastani@uu.nl

Shihan Wang
Utrecht University
Utrecht, Netherlands
s.wang2@uu.nl

ABSTRACT

Reinforcement Learning (RL) shows promise in optimizing task-oriented dialogue policies, but addressing the challenge of reward sparsity remains challenging. Curriculum learning offers an effective solution by strategically training dialogue policies from simple to complex, facilitating a smooth knowledge transition across varied goal complexities. However, these methods typically assume that goal difficulty will increase gradually to adapt to difficult goals over time. In complex environments lacking intermediate goals, attaining smooth knowledge transitions becomes tricky. This paper proposes a novel Bootstrapped Policy Learning (BPL) framework that adaptively tailors a curriculum for each complex goal through goal shaping, which consists of progressively challenging subgoals. Goal shaping comprises goal decomposition and evolution, breaking complex goals into solvable subgoals and progressively increasing subgoal difficulty as the policy improves. BPL harmoniously combines these aspects, enabling smooth knowledge transitions from simple to complex goals, thereby enhancing task-oriented dialogue policy learning efficiency. Our experiments demonstrate the effectiveness of BPL in two complex dialogue environments.

KEYWORDS

Dialogue Policy; Reinforcement Learning; Curriculum Learning; Goal Shaping

ACM Reference Format:
Yangyang Zhao, Mehdi Dastani, and Shihan Wang. 2024. Bootstrapped Policy Learning: Goal Shaping for Efficient Task-oriented Dialogue Policy Learning: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Task-oriented dialogue (ToD) systems aim at assisting users to complete specific tasks (also referred to as goals) with fewer turns, such as making restaurant reservations, booking taxis or movie tickets. Among these, dialogue policy plays a pivotal role by selecting system responses given the dialogue state input [15]. This selection directly affects the success of the dialogue system. Reinforcement learning (RL) is a powerful learning technique for optimizing a task-oriented dialogue policy [14]. However, the natural reward function of dialogue goals presents a considerable challenge for RL-based dialogue policy optimization, as rewards is sparse and requires a prohibitive amount of exploration to reach the goal and receive some learning signals [2, 11, 12].

Curriculum Learning (CL) strategically arranges the learning order of dialogue policies from easy to difficult to alleviate reward sparsity challenges. This ordered learning strategy enables dialogue policies to leverage information or skills gained from achieving simpler goals as a foundation to aid in the accomplishment of more challenging goals (termed knowledge transition) [6, 8, 17]. These methods typically require goal difficulty to increase gradually over time. Training difficult goals directly leads to dialogue policies requiring numerous rounds of interactions to obtain meaningful rewards, thereby diminishing learning efficiency [7, 9].

To this end, this paper proposes a novel framework, Bootstrapped Policy Learning (BPL), which employs goal shaping to dynamically tailor a subgoal curriculum for each complex user goal. This subgoal curriculum comprises a sequence of subgoals that incrementally increase in difficulty, assuring a smooth knowledge transition. Goal shaping involves two key operations: goal decomposition and goal evolution. Specifically, goal decomposition breaks down complex goals into subgoals with solvable maximum difficulty, reducing their complexity. Meanwhile, goal evolution gradually increases the difficulty of subgoals in line with the policy's growing capabilities, ultimately enabling mastery of the entire goal. On the one hand, BPL efficiently navigates the policy's progression from easier to more difficult goals, ensuring a smooth knowledge transition. On the other hand, the customized subgoal curriculum aligns with the policy's evolving abilities, making training more efficient. We constructed experiments on two complex dialogue datasets and verified the effectiveness of our BPL.

2 DIFFICULTY EVALUATION

The measure of successful dialogue depends on accurately identifying all provided information C from the user, correctly responding to all user requests R, and effectively reserving a ticket meeting the specified information. Thus, the complexity of user goal g varies according to the number of attributes present in C and R. Fewer constraints and requests result in fewer agent actions required to complete g, reducing the risk of errors. Based on the defined difficulty of user goals, we present the core ideas behind goal shaping:

• *Goal Decomposition*: reducing the number of attribute in the user goal to reduce its difficulty;

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

• *Goal Evolution*: increasing the number of attribute in the user goal to enhance its difficulty.

3 BOOTSTRAPPED POLICY LEARNING

Bootstrapped Policy Learning is composed of two integral components: *Decomposer* and *Evolver*. *Decomposer* breaks down the user goal into a subgoal with maximal solvable difficulty for goal decomposition. *Evolver* increases the complexity of the user goal for goal evolution, until the dialogue policy masters the entire user goal. Through the interplay of these components, BPL tailors a curriculum aligned with the dialogue policy's capabilities for each goal, thus effectively adapting to complex dialogue environments.

Decomposer 's role is to decompose tricky user goals during RL training phase. Decomposing simple user goals is counterproductive and hampers learning efficiency. Thus, the user goal is considered a tricky one to decompose if it continues to fail after a period of dialog policy training. Goal decomposition consists of three stages:

i) *Boundary state detection* identifies the state nearest to the goal state within a failed dialogue trajectory based on the sampled user goal. The state with the shortest distance d from the goal state is the boundary state. d is determined by the number of mismatched attribute-value pairs: $d = N(s_g) - N(s)$, where $N(s_g)$ and $N(s)$ denote the number of consistent attribute-value pairs of the goal state s_g and the state s.

ii) *Goal Decomposition* divides the current user goal into a corresponding boundary subgoal based on the detected boundary state. Based on attribute-value pairs present in the detected boundary state, the user goal is decomposed into two parts: the boundary subgoal, containing attributes from the boundary state, and the failed subgoal, comprising the remaining attributes in the user goal.

iii) *Goal Substitution* substitutes the current user goal with the boundary subgoal.

Evolver 's role is to increase the complexity of the easy goals that the dialogue policy has already mastered. Thus, the user goal is considered an easy one to evolve if it has been successfully attained. We employ the $Evaluator(D, g)$[1] function to employ whether dialogue D achieves goal g. Goal evolution is initiated when $Evaluator(D, g)$ returns True, and comprises three stages:

i) *Evolutionary segmentation* divides the failed subgoal into an evolved part for subgoal evolution and a retained part for the next iteration. A attribute-value pair within the failed subgoal is randomly designated as the evolved part ready for evolution, while the remaining pairs constitute the retained part.

ii) *Subgoal evolution* merges the evolved part and the current goal into a new goal. The attribute-value pairs from the evolved part are merged into the subgoal.

iii) *Goal Substitution* replaces the original user goal g_i with the evolved new goal.

4 EXPERIMENTAL RESULTS

Experiments are conducted on two datasets with a publicly available user simulator: Taxi Ordering and Multiwoz 2.1 [1, 4, 5]. The former one contains single domain, while Multiwoz is a multi-domain dataset spanning seven domains, all of which are complex dialog environments.

[1]https://github.com/thu-coai/Convlab-2

Table 1: Results of different agents on two complex datasets. The difference between the results of all agent pairs evaluated at the same epoch is statistically significant ($p < 0.01$).

Agent	Taxi			Multiwoz		
	Success	Rewards	Turns	Success	Rewards	Turns
DQN	0.3635	-7.18	21.81	0.1223	-41.91	35.17
SNA-DQN	0.0000	-42.08	26.16	0.0115	-57.02	38.79
SND-DQN	0.0000	-41.62	25.24	0.0153	-56.68	39.02
ACL-DQN	0.5874	13.90	19.92	0.0584	-50.63	37.26
SDPL	0.6318	18.22	18.27	0.0294	-54.72	38.49
VACL	0.5675	12.51	19.13	0.0544	-51.25	37.55
HRL	0.3783	-3.92	21.24	0.2564	-24.56	28.67
SDN	0.6209	17.24	19.29	0.0986	-45.20	35.05
BPL	**0.6972**	**24.75**	**17.99**	**0.3239**	**-14.18**	**28.11**

We compare BPL with RL-based representative algorithm: DQN [3], RL-based algorithms integrating CL: SNA-DQN [10], SND-DQN [10], ACL-DQN [17], SDPL [6], and VACL [16], as well as RL-based algorithms utilizing goal decomposition: HRL [9] and SDN [13]. The main results are presented in Tab.1. Due to the random selection of user goals, DQN often samples difficult user goals, resulting in learning inefficiency. Although SNA-DQN and SND-DQN establish learning sequences based on the number of attributes, this crude assessment of difficulty and inflexible learning have adverse effects. In contrast, VACL and SDPL employ more precise difficulty assessment criteria, demonstrating superior performance on the Taxi dataset. However, for the highly challenging multi-domain dataset Multiwoz, even the simplest user goals prove too difficult for dialogue policies. Consequently, these curriculum learning-based algorithms lose their effectiveness. HRL can reduce goal difficulty by decomposing targets but excels only in multi-domain datasets with easily separable domains, showing little variation in single-domain scenarios where goal separation is challenging. The effectiveness of SDN relies on a wealth of successful dialogues to aid in user goal decomposition, thus its performance is less evident in the sparse successful dialogues of the Multiwoz dataset. Our BPL outperforms other baselines on both challenging datasets, with a more pronounced advantage on the multi-domain Multiwoz dataset. In conclusion, by showingconsistent results across different data sources in both single and multi-domain settings, the BPL framework proves highly effective in various challenging dialogue tasks.

5 CONCLUSION AND FUTURE WORK

This work proposes a novel Bootstrapped Policy Learning (BPL) framework that effectively handles complex dialogue environments, leading to efficient task-oriented dialogue policy learning. This is achieved by dynamically generating progressively challenging subgoal curriculum for each complex goal through goal shaping, involving two key operations: 1) goal decomposition, extracting a solvable boundary subgoal from user goals based on dialogue trajectories, and 2) goal evolution, progressively increasing the difficulty of boundary subgoals until mastery of the entire goal. In the future, our focus will explore the mechanisms for transferring the knowledge acquired from subgoals to new agents.

REFERENCES

[1] Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 5016–5026. https://aclanthology.org/D18-1547/

[2] Wai-Chung Kwan, Hongru Wang, Huimin Wang, and Kam-Fai Wong. 2023. A Survey on Recent Advances and Challenges in Reinforcement Learning Methods for Task-oriented Dialogue Policy Learning. *Int. J. Autom. Comput.* 20, 3 (2023), 318–334. https://doi.org/10.1007/s11633-022-1347-y

[3] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-End Task-Completion Neural Dialogue Systems. In *IJCNLP*, Greg Kondrak and Taro Watanabe (Eds.). Asian Federation of Natural Language Processing, 733–743. https://aclanthology.org/I17-1074/

[4] Xiujun Li, Zachary C. Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. 2016. A User Simulator for Task-Completion Dialogues. *CoRR* abs/1612.05688 (2016). http://arxiv.org/abs/1612.05688

[5] Xiujun Li, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft Dialogue Challenge: Building End-to-End Task-Completion Dialogue Systems. *CoRR* abs/1807.11125 (2018). http://arxiv.org/abs/1807.11125

[6] 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 the Association for Computational Linguistics: ACL/IJCNLP (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 1091–1102. https://doi.org/10.18653/v1/2021.findings-acl.94

[7] Keting Lu, Shiqi Zhang, and Xiaoping Chen. 2019. Goal-Oriented Dialogue Policy Learning from Failures. In *AAAI*. AAAI Press, 2596–2603. https://doi.org/10.1609/aaai.v33i01.33012596

[8] 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 (2020), 181:1–181:50. http://jmlr.org/papers/v21/20-212.html

[9] Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning. In *EMNLP*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 2231–2240. https://doi.org/10.18653/v1/d17-1237

[10] Atsushi Saito. 2018. Curriculum Learning Based on Reward Sparseness for Deep Reinforcement Learning of Task Completion Dialogue Management. In *EMNLP*, Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail S. Burtsev (Eds.). Association for Computational Linguistics, 46–51. https://doi.org/10.18653/v1/w18-5707

[11] Pei-Hao Su, Milica Gasic, and Steve J. Young. 2018. Reward estimation for dialogue policy optimisation. *Comput. Speech Lang.* 51 (2018), 24–43. https://doi.org/10.1016/j.csl.2018.02.003

[12] Ryuichi Takanobu, Runze Liang, and Minlie Huang. 2020. Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition. In *ACL*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 625–638. https://doi.org/10.18653/v1/2020.acl-main.59

[13] Da Tang, Xiujun Li, Jianfeng Gao, Chong Wang, Lihong Li, and Tony Jebara. 2018. Subgoal Discovery for Hierarchical Dialogue Policy Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 2298–2309. https://doi.org/10.18653/v1/d18-1253

[14] Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proc. IEEE* 101, 5 (2013), 1160–1179. https://doi.org/10.1109/JPROC.2012.2225812

[15] Haodi Zhang, Zhichao Zeng, Keting Lu, Kaishun Wu, and Shiqi Zhang. 2022. Efficient Dialog Policy Learning by Reasoning with Contextual Knowledge. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 11667–11675. https://ojs.aaai.org/index.php/AAAI/article/view/21421

[16] Yang Zhao, Hua Qin, Zhenyu Wang, Changxi Zhu, and Shihan Wang. 2022. A Versatile Adaptive Curriculum Learning Framework for Task-oriented Dialogue Policy Learning. In *Findings of the Association for Computational Linguistics: NAACL*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 711–723. https://doi.org/10.18653/v1/2022.findings-naacl.54

[17] Yangyang Zhao, Zhenyu Wang, and Zhenhua Huang. 2021. Automatic Curriculum Learning With Over-repetition Penalty for Dialogue Policy Learning. In *AAAI*. AAAI Press, 14540–14548. https://ojs.aaai.org/index.php/AAAI/article/view/17709