Thesis Report : Generation and Evaluation of Goal-Oriented Dialog with Policy Gradients

Shishir Narayan
Berlin

August 7, 2018

Abstract

1 Introduction

For decades now humans and computers have interacted through Spoken Dialogue Systems(SDS). Here we focus on goal-oriented dialogues where the machine helps to user (a sales agent) to achieve a certain goal. In this case to successfully convert or generate a sales lead though a conversation with a potential customer.

Dialogue Management (DM) [1] is the primary task of the SDS. Provided that contextual information such as the history of the internal states of the dialogue, database hits, API calls, and other domain specific information, which could collectively called *dialogue context*, the DM should essential predict the right next action to utter to the user. Thus the DM has to take the best actions i.e. make good *decisions* based on often times incomplete and highly variant *contexts* or observed states in order to eventually achieve a *reward*. This definition of the problem allows us to cast dialogue management as a sequential decision making problem. First done by Pieraccini et. al [2] who cast the DM problem as a Markov Decision Process (MDP) [3].

2 Background

2.1 Markov Decision Processes

MDPs are a mathematical framework that facilitates the learning of an optimal mapping between situations(or states) and actions. A *policy* is often the name given to this mapping. Policy learning often occurs through the learning of a *action-value function* or a *state-value function* or both. Formally an MDP is defined by a tuple $\{S, A, P, R, \gamma\}$. Here A is a the discrete action space, S is the state space, P is the transition probabilities, R is the reward function, and $\gamma \in [0, 1]$ is called the discount factor that prioritizes short-term rewards. At each time step t, a particular state s_t characterizes the environment. The agent has to now choose an action a_t according to policy, $\pi : S \rightarrow A$. Due to this interaction with the state, it changes to s_{t+1} according to the transition probabilities and this change could lead to a feedback to the agent as the reward, $r_t = R(s_t, a_t, s_{t+1})$. The goal of the agent is to then find a policy which maximizes the expected discounted cumulative reward. Simply, this quantity can be defined as:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \ldots + R_T \tag{1}$$

Where T is the final time-step. To deal with the case of infinite time-steps, we use the concept of *discounting*. Here, the agent tried to select actions so that the sum of the discounted rewards is maximised. The *discount rate* , $0 \leq \gamma \leq 1$ determines the present value of future rewards.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2}$$

Further, (2) can be simplified as successive returns are related to each other.

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} \ldots \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} \ldots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{3}$$

Here G_t is expected discounted cumulative reward and t defines the current time-step. Now we can define the *value* of a state s given a policy π is the total expected return when starting in s and henceforth following π.

$$v_\pi(s) = E_\pi[G_t|S_t = s] = E_\pi\Big[\sum_{t+1}^{T} \gamma^{t+1} R_{t+1} \Big| S_t = s\Big],\ for\ all\ s \in S, \tag{4}$$

where $E_\pi[.]$ is the expected value of a given state if the agent follows policy π and t is any time-step. v_π is known as the *state-value function* for *policy* π. Likewise, we can define the value of taking an action a in state s given a policy π, as the expected return starting from s, taking action a and thereafter following policy π:

$$q_\pi(s,a) = E_\pi[G_t|S_t = s,\ A_t = a] = E_\pi\Big[\sum_{t+1}^{T} \gamma^{t+1} R_{t+1} \Big| S_t = s,\ A_t = a\Big] \tag{5}$$

We call the $q_\pi(s,a)$the *action-value function* for *policy* π

2.2 Optimality

In essence solving the Reinforcement Learning task aims to find a policy that selects actions in a way that maximises the future rewards. In the case of finite MDPs, a policy π is said to be better than or equal to a policy π' if it's expected return is greater than or equal to that of π' for all states. So $\pi \geq \pi'$ if and only if $v_\pi(s) \geq v_{\pi'}(s)$ for all $s \in S$. This implies that there is always at least one policy that is better than or equal to all other policies. This is denoted by π_*, and it's corresponding state-value function as v_*. and is shared by all optimal policies. Thus

$$v_*(s) = \max_\pi v_\pi(s),\ for\ all\ s \in S \tag{6}$$

The *optimal action-value function* is also shared between optimal policies and is denoted by q_*

$$q_*(s,a) = \max_\pi q_\pi(s,a),\ for\ all\ s \in S \tag{7}$$

References

[1] S. Young, "Probabilistic methods in spoken dialogue systems," *Philosophical Transactions of the Royal Society (Series A*, vol. 358, pp. 1389–1402, 1999.

[2] E. Levin, R. Pieraccini, and W. Eckert, "Learning dialogue strategies within the markov decision process framework," pp. 72 – 79, 01 1998.

[3] R. Bellman, "A markovian decision process," vol. 6, p. 15, 04 1957.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning : An Introduction.* MIT Press, 1998.

References

[1] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu and Tat-Seng Chua and Luigi R. Bedin, Neural Collaborative Filtering, CoRR, abs/1708.05031 (2017).

[2] Richard S. Sutton and Andrew G., Reinforcement Learning : An Introduction, MIT Press, (1998).

[3] Antoine Bordes and Jason Weston, Learning End-to-End Goal-Oriented Dialog, CoRR,abs/1605.07683 (2016), http://arxiv.org/abs/1605.07683

[4] Ilya Sutskever and Oriol Vinyals and Quoc V. Le, Sequence to Sequence Learning with Neural Networks, CoRR,abs/1409.3215 (2014), http://arxiv.org/abs/1409.3215

[5] Li Zhou and Kevin Small and Oleg Rokhlenko and Charles Elkan, End-to-End Offline Goal-Oriented Dialog Policy Learning via Policy Gradient, CoRR,abs/1712.02838 (2017), http://arxiv.org/abs/1712.02838 1711.01731

[6] Salimans, Tim; Ho, Jonathan; Chen, Xi; Sidor, Szymon; Sutskever, Ilya, Evolution Strategies as a Scalable Alternative to Reinforcement Learning, arXiv:1703.03864 (2017), http://arxiv.org/abs/1703.03864

[7] Ronald J Williams, Simple Statistical Gradient-Following Algorithms for. Connectionist Reinforcement Learning, Machine Learning, 8, pp. 229-256 (1992), http://www-anw.cs.umass.edu/ barto/courses/cs687/williams92simple.pdf

[8] T. Zhao and M. Eskenazi, Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning, In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 110, Los Angeles, September 2016. Association for Computational Lin- guistics.

[9] H. Cuayhuitl, S. Keizer, and O. Lemon, dialogue management via deep reinforcement learning, arxiv.org, 2015

[10] T.-H. Wen, D. Vandyke, N. Mrksi c, M. Gasic, L. M. Rojas Barahona, P.-H. Su, S. Ultes, and S. Young. A, network-based end-to-end trainable task-oriented dialogue system, Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 438449, Valencia, Spain, April 2017. Association for Computational Linguistics

[11] Rumelhart, D. E., Hinton, G. E., Williams, R. J., Learning representations by back-propagating errors, Cognitive modeling, 5(3):1.

[12] J. D. Williams and G. Zweig, End-to-end lstm-based dialog control optimized with supervised and reinforcement learning, Cognitive modeling, 5(3):1.

[13] X. Li, Y.-N. Chen, L. Li, and J. Gao, End-to-end task- completion neural dialogue systems, arXiv preprint arXiv:1703.01008, 2017.

[14] Csky, Richrd, Deep Learning Based Chatbot Models, 10.13140/RG.2.2.21857.40801.

[15] Hongshen Chen and Xiaorui Liu and Dawei Yin and Jiliang Tang, A Survey on Dialogue Systems: Recent Advances and New Frontiers, CoRR,abs/1711.01731 (2017), http://arxiv.org/abs/1711.01731

[16] Anu Venkatesh and Chandra Khatri and Ashwin Ram et. al, On Evaluating and Comparing Conversational Agents, CoRR,abs/1801.03625 (2018), http://arxiv.org/abs/1801.03625

[17] Sebastian Mller et. al., MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations , Deutsche Telekom Labs, https://pdfs.semanticscholar.org/bd47/552316528591dc1feae50675fd0e7be9c289.pdf

[18] Kamm, C., User interfaces for voice applications, (1995), Proceedings of the National Academy of Sciences of the United States of America, 92(22), 1003110037.

[19] Jiwei Li and Will Monroe and Tianlin Shi et. al., Adversarial Learning for Neural Dialogue Generation, CoRR,abs/1701.06547 (2017), http://arxiv.org/abs/1701.06547

[20] Joshi, C. K., Mi, F., and Faltings, B., Personalization in goal-oriented dialog, CoRR,abs/1706.07503 (2017), http://arxiv.org/abs/1706.07503

[21] Zhao, T., Lu, A., Lee, K., and Eskenazi, M, Generative encoder- decoder models for task-oriented spoken dialog systems with chatting capability, CoRR,abs/1706.08476 (2017), http://arxiv.org/abs/1706.08476

[22] Li, J., Miller, A. H., Chopra, S., Ranzato, M., and Weston, J.., Dialogue learning with human-in-the-loop, CoRR,abs/1611.09823 (2017), http://arxiv.org/abs/1611.09823