Challenges in Explainability and Knowledge Extraction

Daiki Kimura **Tsunehiko Tanaka** [†*] **Michiaki Tatsubori** **Asim Munawar**
IBM Research, [†]Waseda University
daiki@jp.ibm.com

Abstract

Since text-based games are one of the most challenging problems and require to have an understanding of language and decision-making in a complex environment, there are many studies about applying deep reinforcement learning methods. However, these deep models are often black-box which means the human operator cannot understand the trained rules. And the models use external knowledge such as common-sense knowledge from well-designed data. In this paper, we explain some challenges in text-based games.

1 Introduction

Text-based games (Côté et al., 2018; Hausknecht et al., 2019; Keerthiram Murugesan and Campbell, 2021) are appropriate test-bed for tackling various challenges in reinforcement learning and natural language processing. There are some methods which solve various challenges. LSTM-DQN (Narasimhan et al., 2015) is a study on an LSTM-based encoder for feature extraction from observation and Q-learning for action policy. LSTM-DQN++ (Yuan et al., 2018) extended the exploration and LSTM-DRQN was proposed for adding memory units in the action scorer. KG-DQN (Ammanabrolu and Riedl, 2019) and GATA (Adhikari et al., 2020) extended the language understanding. LeDeepChef (Adolphs and Hofmann, 2020) used recurrent feature extraction along with the A2C (Mnih et al., 2016). CREST (Chaudhury et al., 2020) was proposed for pruning observation information. These methods tackled to solve the game by purely deep methods.

Recently, some studies (Kimura et al., 2021b; Chaudhury et al., 2021; Kimura et al., 2021a) introduced symbolic reasoning to train logical rules in a recent neuro-symbolic framework called the Logical Neural Network (LNN) (Riegel et al., 2020) has been proposed to simultaneously provide key properties of both the neural network (learning) and the symbolic logic (reasoning). FOL-LNN (Kimura et al., 2021b) introduced training of first-order logic for increasing the training speed and interpretability. SLATE (Chaudhury et al., 2021) proposed interpretable action policy rules from symbolic abstractions of textual observations for improved generalization. LOA (Kimura et al., 2021a) is an open-source code for neuro-symbolic reinforcement learning method. However, we still cannot interact to edit the logical rules.

On the other hand, there are some studies (Keerthiram Murugesan and Campbell, 2021; Tanaka et al., 2022) about introducing the common-sense knowledge as external knowledge. TWC (Keerthiram Murugesan and Campbell, 2021) is proposed as a test-bed game for using the common-sense knowledge in text-based game. We (Tanaka et al., 2022) proposed a method just uses common-sense knowledge from scene graph dataset. However, we still cannot utilize the external knowledge from large data.

In this paper, we explain challenges in explainability and knowledge extraction.

2 Challenges

2.1 Explainability

We shows a web-based interactive UI for playing the text-based game, and visualizes the trained logical rules in LNN (Kimura et al., 2021a). They recently shared a new demo UI for easily understanding at the github repository [1]. However, it is just showing the trained knowledge.

We are thinking that there is a challenge to realize an interactive interface with a human operator for solving text-based games informatively. For example, the human operator can see the reasons (such as why the agent takes "put a used tissue into

[*]Work done during internship at IBM Research

[1]https://github.com/IBM/nesa-demo

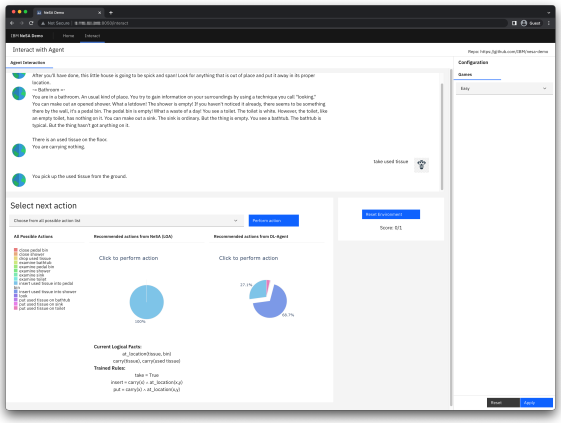

Figure 1: Demo for Neuro-Symbolic Agent

pedal bin" action) for trained rules, and interact to edit the acquired rules during the execution of reinforcement learning. These functions will give hints on how to operate the agent in the environment, and understanding the agent behavior.

2.2 Knowledge Extraction

We introduced scene graph datasets as external knowledge with abstraction levels close to actual tasks (Tanaka et al., 2022). However, it missed some valid common-senses and failed to deal noise.

We believe that a challenge is to extract noiseless common sense by understanding meaning of word, and use the extracted common-sense have to be accurately linked to tasks. Specifically, since scene graph dataset has common-sense knowledge that is not relevant to tasks, the model should prune in the pre-processing of encoding common-sense. It is also necessary clearly indicate which common-sense should be used for which task because the amount of common-sense is enormous.

3 Conclusion

We discuss the challenges in explainability and knowledge extraction for text-based games.

References

Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and William L. Hamilton. 2020. Learning dynamic belief graphs to generalize on text-based games. In *NeurIPS*.

Leonard Adolphs and Thomas Hofmann. 2020. Ledeepchef deep reinforcement learning agent for families of text-based games. In *AAAI*, volume 34, pages 7342–7349.

Prithviraj Ammanabrolu and Mark Riedl. 2019. Playing text-adventure games with graph-based deep reinforcement learning. In *North American Chapter of the Association for Computational Linguistics*, pages 3557–3565.

Subhajit Chaudhury, Daiki Kimura, Kartik Talamadupula, Michiaki Tatsubori, Asim Munawar, and Ryuki Tachibana. 2020. Bootstrapped q-learning with context relevant observation pruning to generalize in text-based games. In *EMNLP*, pages 3002–3008.

Subhajit Chaudhury, Prithviraj Sen, Masaki Ono, Daiki Kimura, Michiaki Tatsubori, and Asim Munawar. 2021. Neuro-symbolic approaches for text-based policy learning. In *EMNLP*, pages 3073–3078.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew J. Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. Textworld: A learning environment for text-based games. In *IJCAI workshop on Computer Games Workshop*, pages 41–75.

Matthew Hausknecht, Prithviraj Ammanabrolu, Côté Marc-Alexandre, and Yuan Xingdi. 2019. Interactive fiction games: A colossal adventure. *CoRR*, abs/1909.05398.

Pavan Kapanipathi Pushkar Shukla Sadhana Kumaravel Gerald Tesauro Kartik Talamadupula Mrinmaya Sachan Keerthiram Murugesan, Mattia Atzeni and Murray Campbell. 2021. Text-based RL Agents with Commonsense Knowledge: New Challenges, Environments and Baselines. In *AAAI*.

Daiki Kimura, Subhajit Chaudhury, Masaki Ono, Michiaki Tatsubori, Don Joven Agravante, Asim Munawar, Akifumi Wachi, Ryosuke Kohita, and Alexander Gray. 2021a. LOA: Logical optimal actions for text-based interaction games. In *ACL*, pages 227–231.

Daiki Kimura, Masaki Ono, Subhajit Chaudhury, Ryosuke Kohita, Akifumi Wachi, Don Joven Agravante, Michiaki Tatsubori, Asim Munawar, and Alexander Gray. 2021b. Neuro-symbolic reinforcement learning with first-order logic. In *EMNLP*, pages 3505–3511.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937.

Karthik Narasimhan, Tejas D. Kulkarni, and Regina Barzilay. 2015. Language understanding for text-based games using deep reinforcement learning. In *EMNLP*, pages 1–11.

Ryan Riegel, Alexander Gray, Francois Luus, Naweed Khan, Ndivhuwo Makondo, Ismail Yunus Akhalwaya, Haifeng Qian, Ronald Fagin, Francisco Barahona, Udit Sharma, Shajith Ikbal, Hima Karanam, Sumit Neelam, Ankita Likhyani, and Santosh Srivastava. 2020. Logical neural networks.

Tsunehiko Tanaka, Daiki Kimura, and Michiaki Tatsubori. 2022. Commonsense knowledge from scene graphs for textual environments. *AAAI Workshop on Reinforcement Learning in Games*.

Xingdi Yuan, Marc-Alexandre Côté, Alessandro Sordoni, Romain Laroche, Remi Tachet des Combes, Matthew J. Hausknecht, and Adam Trischler. 2018. Counting to explore and generalize in text-based games. *CoRR*, abs/1806.11525.