強化学習エージェントにおける Commonsense Knowledge の調査

A Survey of Commonsense Knowledge in Reinforcement Learning Agents

村上 遼太郎
Ryotaro Murakami

岡山大学 太田研究室
Ohta Laboratory, Okayama University

概要 人間らしい思考をもつマシンの構築は人工知能における大きな課題のひとつである．特に，エージェントに人間の常識（Commonsense Knowledge, CK）を持たせることは，自然な対話や環境に適応した行動を実現するために不可欠である．本稿では，近年行われているエージェントにおける CK 獲得の研究とその学習手法である Reinforcement Learning from Human Feedback (RLHF) および Deep Q-Network (DQN) について調査し，今後の研究展開について検討し方針をまとめる．

1 はじめに

人間の常識（Commonsense Knowledge, CK）は日常生活における自然な対話や環境に応じた行動の基盤となるものであり，文脈や状況に応じた柔軟な対応を可能にする．人工知能の研究，特に自然言語処理の分野においてエージェントに CK を組み込み，人間らしい思考を与えることは大きな課題のひとつである．

CK は日常的な知識や経験によって得られるため，明確に定義されたルールやデータセットでは表現しきれない．そのため，エージェントが CK を獲得するには単純な大規模データの学習では限界がある．そこで，近年では特に強化学習（Reinforcement Learning, RL）を用いたアプローチが注目されている．RL はエージェントが環境を通じて経験を積むことにより最適な行動を学習する手法であり，人間らしさを獲得することが期待されている．特に，ChatGPT のような Large Language Model (LLM) を利用したチャットボットは，Reinforcement Learning from Human Feedback (RLHF) という強化学習を活用して人とのコミュニケーションにおける CK の獲得を目指している．

本稿では，現在使用されている LLM がどのような CK を獲得できるのか，また CK 獲得のための強化学習手法である RLHF と Deep Q-Network (DQN) を利用した手法について調査した結果をまとめ，今後の課題と私の研究方針について議論する．

2 LLM における CK

人間らしい会話を実現するためには，エージェントが人間の常識を学習する必要がある．そもそも常識とは Wikipedia によると「社会的に当たり前と思われる行為，その他物事のこと」とされている．

では，機械はどのようにして人間の当たり前の知識や行為を学習すればよいのだろうか．人工知能の父とも呼ばれるマーヴィン・ミンスキー氏は常識について「常識というのは単純なものではない．逆に常識は，苦しみの末に身についた，たくさんの実用的な考えからなる巨大な社会である．」と述べており，大部分は単純な法則や規則性のような原理に基づいていないことを指摘している [4]．つまり，自然言語処理における CK の獲得には，テキストデータのパターンを学習するだけでは不十分であり，エージェントが文章の意味を理解できないことが予想される．実際に李ら [6] は既存の LLM がどの程度 CK を有しているのか調査した．2800 億パラメータを持つトランスフォーマーモデルを用いて CK 性能を測る 4 つのベンチマークを評価した．評価結果より，モデルサイズの増加によりある程度の性能の向上は見られるものの，人間のレベルに達するには現在のモデルよりも遥かに大きなデータが必要であると結論付けた．

しかしながら，大規模なテキストデータの学習には莫大な時間とコストがかかるため現実的ではない．したがって，LLM が効率的に CK を獲得するためには強化学習による経験の積み重ねを行うことが有効である．

3 強化学習による CK の獲得

本節では CK 獲得のために使用されている強化学習の手法について紹介する．

3.1 RLHF

RLHF は人間のフィードバックを利用した強化学習手法 [3] である．RLHF は図 1 に示す 3 つの学習プロセスからなる．Step1 では教師ありファインチューニング（Supervised Fine-tuning, SFT）を使用する．学習に用いるプロンプトとそれに対応した望ましい出力を人が用意し言語モデルをファインチューニングする．Step2 では，報酬モデル（Reward Model, RM）をトレーニングする．Step1 で作成したモデルを利用し，一つのプロンプトに対する複数の応答を用意し，好ましい順に人間がランク付けを行う．報酬モデルをこのように学習することで，似たようなプロンプトが与えられた時により望ましいと評価された応答に近い応答文がより報酬を得ることができる．Step3 では報酬モデルに

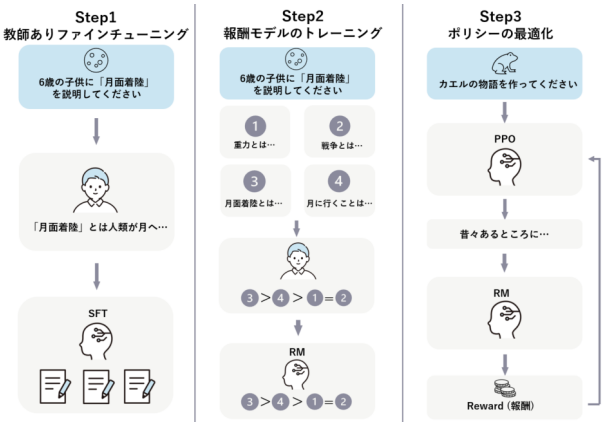

図 1: RLHF の概要

対してポリシーを最適化する．強化学習におけるポリシーとは，現在の状態で次にエージェントがどういう行動をとるべきかを決定する戦略を意味する．OpenAI の RLHF では，近接ポリシー最適化（Proximal Policy Optimization, PPO）[2] というアルゴリズムを用いている．Step3 の学習の流れとしては，まずデータセットから新たなプロンプトを抽出する．次に，その時点でのポリシーに従って抽出したプロンプトから応答文を生成し，生成された応答文を報酬モデルに与えて報酬を計算する．得られた報酬を使って現在採用されているポリシーをより高い報酬が得られると期待できるポリシーへ変更する．以上のプロセスを繰り返すことで人間らしい回答を学習することができる．

3.2 DQN を利用した CK 獲得手法

Q 学習とは，ある状態である行動をとった際にその後の収益の期待値（Q 値）を計算し，この Q 値を更新しながら学習を進める手法で，DQN は Q 学習にニューラルネットワークを導入した強化学習手法である．Q 学習においては式 (1) で Q 値を求める

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max Q(s',a')\right) \tag{1}$$

ここで s, a, s', a', r はそれぞれ現状態，現状態において実行された行動，次状態，次に実行される行動，状態 s で行動 a を実行したときの報酬を表している．また，学習率は α，割引率は γ で表される．DQN の学習では次状態の Q 値やそこで獲得する報酬に基づき Q 値を計算し，ニューラルネットワークが出力した Q 値との損失関数を計算し，誤差逆伝搬を行う．

学習の例として，Joshi らの研究 [1] で提案された `ScriptWorld` というフレームワークを用いた DQN を紹介する．`ScriptWorld` の概要を図 2 に示す．まず時系列的なイベントの説明をまとめた複数の ESD（Event Sequence Description）に含まれる，意味的に類似したイベントをクラスタリングする．次に作成したクラスタをノードとし，イベントの順序に基づきノードを

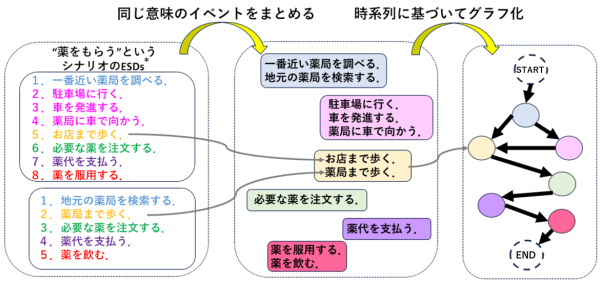

図 2: ScriptWorld における ESDs を用いた環境グラフ作成のプロセス

連結することで有向グラフを作成する．これを環境グラフと呼ぶ．強化学習エージェントはグラフのノードを状態として観測し，ノードに隣接した正しい選択肢を選び報酬を得ることで学習を進める．久保ら [5] は `ScriptWorld` の環境に DQN を適用し，CK を学習する手法を提案した．実験結果より Q 学習と比較して性能の向上がみられ，入力情報の汎化による CK の獲得に成功した．

4 まとめと今後の研究方針

本報告では強化学習エージェントにおける CK 獲得のための学習手法である RLHF と DQN を紹介した．RLHF は人間らしさを AI モデルに反映する点では優れているが，意思決定において人間と同レベルの意思決定能力を実現するには膨大なデータが必要である．したがって，今後は人間の介在を必要としない強化学習による CK 獲得が課題となる．具体的には，`ScriptWorld` における ESD のクラスタリングを自動化することで，強化学習の効率化を目指す．

参考文献

[1] Abhinav Joshi *et al.*, "ScriptWorld: Text Based Environment for Learning Procedural Knowledge," Proc. of the 32nd International Joint Conference on Artificial Intelligence, pp. 5095–5103, 2010.
[2] John Schulman *et al.*, "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347, 2017.
[3] Long Ouyang *et al.*, "Training Language Models to Follow Instructions with Human Feedback," Advances in Neural Information Processing Systems, NeurIPS, vol. 35, pp. 27730–27744, 2022.
[4] マーヴィン・ミンスキー著．安西祐一郎訳．心の社会．産業図書，2000，p. 14.
[5] Ryota Kubo *et al.*, "Reward Design for Deep Reinforcement Learning Towards Imparting Commonsense Knowledge in Text-based Scenario," Proc. of the 16th International Conference on Agents and Artificial Intelligence, pp. 1213–1220, 2024.
[6] Xiang Lorraine Li *et al.*, "A Systematic Investigation of Commonsense Knowledge in Large Language Models," Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 11838–11855, 2022.