2Q-02

強化学習に基づく自動交渉戦略のための事前学習モデル

小林 裕二 †　　藤田 桂英 ‡

† 東京農工大学　工学部　知能情報システム工学科　　‡ 東京農工大学大学院　グローバルイノベーション研究院

1　はじめに

自動交渉は，エージェント同士が異なる目的や利害関係を持ちながら自動で交渉を行う技術であり，競合が生じている中でもエージェントが自律性を保ちながら合意形成することができる．近年，二者間複数論点交渉問題における交渉戦略の獲得のために強化学習が適用されており，その汎用性から注目を集めている．

先行研究において，Higa ら [1] は自動交渉戦略を獲得する汎用的な End-to-End の強化学習フレームワークを提案した．しかし，適切な交渉戦略を学習するためには膨大な訓練を対戦相手や交渉ドメインごとに行う必要がある．

本研究では，より短時間かつ効率的な訓練を実現するため，自然言語処理の分野で注目されている事前学習モデルとファインチューニングを，End-to-End の強化学習フレームワークに適用するアプローチを提案する．そして，事前学習モデルとファインチューニング適用後のモデルをそれぞれ評価する．

2　問題設定

本研究では，2 つのエージェント同士で複数の論点について交渉を行う二者間複数論点交渉問題を扱う．交渉プロトコルについては，二者間交渉で広く利用されている Alternating Offers Protool[2] を用いる．このプロトコルは，2 つのエージェントが交互に行動する設定となっており，交渉中のエージェントは提案を受け入れるか，拒否して新たな提案を送るかを選択することができる．交渉終了後はスコアとして，ある合意案に対するエージェントの選好度を表す効用値を，各エージェントそれぞれが獲得する．これを最大化することが本問題における交渉エージェントの目的である．

A Pre-trained Model for Automated Negotiation Strategies based on Reinforcement Learning

†Department of Electrical Engineering and Computer Science, Faculty of Engineering, Tokyo University of Agriculture and Technology

‡Institute of Global Innovation Research, Tokyo University of Agriculture and Technology

3　交渉戦略のための End-to-End 強化学習フレームワーク (VeNAS)

Higa ら [1] は自動交渉戦略を獲得する合意案候補 (Bid) ベースの汎用的な End-to-End の強化学習フレームワークである PPO-VeNAS を提案した．PPO-VeNAS の入力は相手から受け取った Bid，出力は次に相手に送るメッセージ (受諾もしくは提案 Bid) である．また，強化学習のフレームワークは PPO[3] を用いている．過去の国際自動交渉エージェント競技会 (ANAC) 優勝エージェントを含む既存の交渉エージェントで作成したベースラインよりも同等かそれ以上の効用値を獲得できることを示した．しかし，膨大な訓練を対戦相手や交渉ドメインごとに行う必要があるという問題点も存在する．

4　提案手法

提案手法の概観を図 1に示す．本提案手法では PPO-VeNAS を用いる．PPO-VeNAS は，ある 1 つのドメインに対して，ある 1 つの対戦相手と交渉を繰り返すことで，その対戦相手に対する交渉戦略を学習する．一方，本研究では，PPO-VeNAS をある 1 つのドメインに対して，複数の対戦相手をランダムに変えながら交渉対戦を繰り返すことで，複数の対戦相手への交渉戦略を学習するようにする．その際，学習後のモデルを事前学習モデル (Pre-trained model) とする．

事前学習モデル作成後，1 つ 1 つの対戦相手に特化するために，事前学習モデルをファインチューニングをする．具体的には，1 つの対戦相手のみと交渉対戦を繰り返すことで，その対戦相手に対する交渉戦略をチューニングする．これにより，既存手法よりもチューニングのための訓練時間を短くすることが目的である．

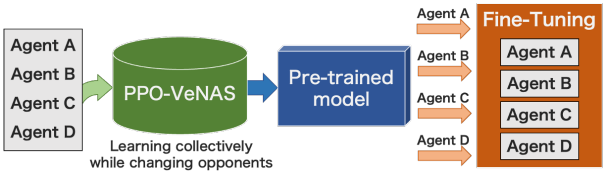

図 1: 本研究の提案手法の概観

5 実験

5.1 実験設定

事前学習モデル及びファインチューニング適用後のモデルを評価するため，交渉シミュレーション実験を行った．実験で用いる交渉ドメインは，Takahashi らの論文 [4] で用いられていた 5 種類のドメインとした．また，制限時間は 40 ラウンドとした．評価する際の効用値は，学習エージェントが先攻の場合と後攻の場合でそれぞれ 100 回対戦相手と交渉を行い，それらを平均した値とした．そして，ベースラインは既存手法である PPO-VeNAS を用い，学習時のステップ数は 50 万ステップとした．

本研究では 2 つの実験を行った．実験 1 では，競技会等で時間依存の交渉戦略が大半であることから，対戦相手を時間依存エージェントの 7 種類とした．実験 2 では，対戦相手として実験 1 で用いたエージェントに加え，行動依存戦略のエージェント及び ANAC 優勝エージェントも加えた 13 種類とした．事前学習モデル作成の際のステップ数はそれぞれ 350 万, 650 万ステップとし，ファインチューニングにおけるステップ数はどちらも 5 万ステップとした．各実験ではベースラインと事前学習モデル，ファインチューニング適用後のモデルを比較性能評価する．

5.2 実験結果

各手法 10 回ずつ学習し，その後交渉シミュレーションによって評価を行い，各ドメインごとに，エージェントの獲得効用値の平均が最大のモデルを用いて最終的な効用値を算出した．実験 1 の性能結果を表 1 に，実験 2 の性能評価を表 2 に示す．

まず, 時間依存戦略のみが対戦相手の場合でも, ANAC 等を含んだ対戦相手の場合でも事前学習モデルの方がほとんどの場合効用値が高いことがわかる．これは，学習エージェントが複数のエージェント戦略を学習することにより汎化性能が向上したことが理由だと考える．

さらに，ファインチューニング適用後のモデルは 1 つを除いた全てのドメインで 3 つのモデルの中で性能が 1 番高いため，ファインチューニングの有用性も示されたと考える．また，事前学習モデルを用いることで，5 万ステップの短時間のファインチューニングで高い性能を獲得している．

表 1: 実験 1 の性能結果

	Baseline	Pre-Trained	Fine-Tuning
Laptop	0.911	**1.000**	**1.000**
ItexvsCypress	0.739	**0.763**	**0.770**
IS_BT_Acquisition	**0.929**	0.926	**0.930**
Grocery	0.924	**0.960**	**0.960**
thompson	0.806	**0.822**	**0.824**

ドメインごとに 1 番高い数値は赤太文字, 2 番目は黒太文字

表 2: 実験 2 の性能結果

	Baseline	Pre-Trained	Fine-Tuning
Laptop	0.878	**0.891**	**0.949**
ItexvsCypress	0.712	**0.758**	**0.800**
IS_BT_Acquisition	0.885	**0.916**	**0.925**
Grocery	0.865	**0.910**	**0.948**
thompson	**0.727**	0.669	**0.697**

ドメインごとに 1 番高い数値は赤太文字, 2 番目は黒太文字

6 まとめ

本稿では，先行研究の膨大な訓練が必要という課題に対し，事前学習モデルの構築とファインチューニングの手法を適用するアプローチを提案した．そして評価実験により，提案手法の有用性を示した．今後の課題としては，ファインチューニングの訓練時間の調整及び実用化に向けてファインチューニングの処理をリアルタイムで学習できるアプローチの提案が挙げられる．

参考文献

[1] Ryota Higa, Katsuhide Fujita, Toki Takahashi, Takumu Shimizu, and Shinji Nakadai. Reward-based negotiating agent strategies. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, number 10, pages 11569–11577, 2023.

[2] John F. Nash Jr. The bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 155–162, 1950.

[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[4] Toki Takahashi, Ryota Higa, Katsuhide Fujita, and Shinji Nakadai. Venas: Versatile negotiating agent strategy via deep reinforcement learning(student abstract). In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, number 11, pages 13065–13066, 2022.