7T-04

深層強化学習を用いた自動交渉における効果的な受入戦略

松尾 飛我 †　　藤田 桂英 ‡

† 東京農工大学　工学部　情報工学科　　‡ 東京農工大学大学院　工学研究院　先端情報科学部門

1　はじめに

近年，マルチエージェントシステムの研究において，エージェントの自律性を保ちつつ競合を解消し合意形成を行うことができる手段として自動交渉が注目されている．自動交渉では主に，相手にどのような提案を送るかを検討する提案戦略と，相手からの提案を受け入れるかを検討する受入戦略の二種類をエージェントに組み込んで交渉を行う．交渉問題の中で最も基本的な二者間複数論点交渉問題では，相手のエージェントと交互に提案を送りあい，相手の提案を受け入れた時点で交渉が終了となる．そのため，エージェントの受入戦略は，自身の効用を高める点で重要である．

本研究では，より高い効用値を得ることを目的とし，Deep Q-Network を用いて相手からの提案を受け入れるかを判断する自動交渉のための深層強化学習フレームワークを新たに提案する．また，様々な報酬関数に対して比較実験を行い，より高い個人効用を得られるものを明らかにする．提案手法を用いて学習を行ったエージェントと既存手法 [1] を用いて学習を行ったエージェントを，他の交渉エージェントとの交渉シミュレーション実験によって効用値を比較することで，提案手法の有効性を示す．

2　問題設定

本研究で扱う二者間複数論点交渉問題では，2 つのエージェントが共通のドメイン上で交渉を行うことを考える．交渉ドメインは，n 個の論点 $I_1, I_2, \ldots, I_n$ と，各論点 I_i における k_i 個の選択肢 $v_1^i, v_2^i, \ldots, v_{k_i}^i$ から構成される．交渉中に提案される合意案候補 (Bid) は各論点から選択肢を一つずつ選んだもので，$\omega = \left[v_{x_1}^1, v_{x_2}^2, \ldots, v_{x_n}^n\right]$ で表される．

交渉を行うエージェントには，それぞれ固有の効用関数が設定される．効用関数は，論点 I_i に対する重み w_i と，各選択肢 v_x^i に対する評価値 $eval(v_x^i)$ から構成される効用値を求める関数である．ただし，論点の重み w_i は $\sum_{i=1}^{n} w_i = 1$ かつ $w_i \geq 0$ を満たし，評価値は $eval(v_x^i) \geq 0$ を満たす．ある合意案候補 ω に対する効用関数 $U(\omega)$ は次の式 (1) で表される．

$$U(\omega) = \sum_{i=1}^{n} w_i \times \frac{eval(v_{x_i}^i)}{\max_j eval(v_j^i)} \tag{1}$$

効用関数 $U(\omega)$ で得られる値を合意案候補 ω の効用値と呼び，0 以上 1 以下の実数として表される．

本研究で扱う交渉では，二者間交渉で広く利用されている Alternating Offers Protocol[2] を用いる．はじめに，一方のエージェントが相手に合意案候補を提案 (Offer) する．他方のエージェントは，この Offer が受け入れられない場合新たな提案を送り，これを制限時間まで交互に繰り返す．ある Offer を相手が受け入れる (Accept) と交渉は終了し，それぞれの効用関数で合意案を評価した効用値を受け取る．制限時間内に合意できなかった場合，獲得効用は 0 となる．

3　提案手法

深層強化学習に用いる環境として，環境の構成要素である入力・行動・報酬について提案する．

入力

時間 t における入力要素を，次の式 (2) に示す．

$$\left[U_A\left(\omega_{B\to A}^{t-1}\right), \frac{t}{T}\right] \tag{2}$$

ただし $U_A(\omega)$ は自身の効用関数による合意案候補 ω の効用値，$\omega_{B\to A}^{t-1}$ は相手が直前に提案した合意案候補，T は制限時間である．従って，式 (2) は相手の提案による効用値，正規化した交渉時間の 2 種を入力としている．

行動

学習するエージェントが取る行動は,「Accept」または「Offer」のどちらかとする．Accept は相手の直前の提案を受け入れ，その提案による効用値を得て交渉を終了するもの，Offer は相手からの提案を拒否し，自分が相手に提案を行うものである．このとき，自分の提案戦略には既存手法である AgentK[3] を用いる．

報酬

報酬関数を，(A) 合意時の報酬，(B) 交渉継続時の報酬，(C) 交渉失敗時のペナルティの 3 種に分けて独立に考える．

(A) 合意時の報酬は，相手からの提案を Accept した場

Effective Acceptance Strategy in Automated Negotiation using Deep Reinforcement Learning

†Department of Computer and Information Sciences, Faculty of Engineering, Tokyo University of Agriculture and Technology

‡Division of Advanced Information Technology and Computer Science, Institute of Engineering, Tokyo University of Agriculture and Technology

合に得る報酬であり，次の式 (3) に示す 3 パターンを提案する．

$$r_a^{accept} = U_A\left(\omega_{B\to A}^{t-1}\right) \tag{3-a}$$

$$r_b^{accept} = \frac{\tanh\left\{5\left(U_A\left(\omega_{B\to A}^{t-1}\right) - 0.5\right)\right\} + 1}{2} \tag{3-b}$$

$$r_c^{accept} = \tanh\left\{5\left(U_A\left(\omega_{B\to A}^{t-1}\right) - 0.5\right)\right\} \tag{3-c}$$

(B) 交渉継続時の報酬は，相手からの提案を拒否し Offer を行った場合に得る報酬であり，次の式 (4) に示す 3 パターンを提案する．

$$r_a^{offer} = 0 \tag{4-a}$$

$$r_b^{offer} = \frac{1 - U_A\left(\omega_{B\to A}^{t-1}\right)}{100} \tag{4-b}$$

$$r_c^{offer} = \frac{1 - average}{100} \tag{4-c}$$

このとき，式 (4-c) の *average* は相手から受け取った提案の効用値の過去 10 回分の平均である．
(C) 交渉失敗時のペナルティは，交渉の制限時間を過ぎても交渉が合意に至らなかった場合に得るペナルティであり，−1，−0.5，0 の 3 パターンを提案する．

4 実験

提案した環境を用いて学習を行い，その性能を評価するために交渉シミュレーション実験を行う．学習の環境は OpenAI Gym を用いて作成し，実験には自動交渉プラットフォーム NegMAS を用いる．

学習には Deep Q-Network を用い，3 章で述べた入力・行動・報酬を設定する．このとき，報酬は (A)(B)(C) の各 1 つずつの組み合わせについてそれぞれ学習を行う．学習に使用する交渉相手のエージェントとして，Gahboninho というエージェントと，boulware エージェントの 2 種を用いる．Gahboninho は，制限時間直前まで自身の効用値が高い提案しかせず，その後譲歩する戦略を持っているエージェントであり，boulware は時間 t における自身の効用値が次の式 (5) に従ったものになるような提案をするエージェントである．

$$U_{target}(t) = 1 - \left(\frac{t}{T}\right)^{\frac{1}{0.25}} \tag{5}$$

また，ドメインは EnglandZimbabwe ドメインを用いる．報酬と相手エージェントの計 54 通りの組み合わせについて，それぞれ独立に 10 回の学習を行う．

実験では，1 回の学習につき交渉を 100 回，組み合わせ 1 つごとに合計 1000 回の交渉を行い，得た効用値の平均や分散を比較する．実験に用いるドメインは party ドメイン，相手エージェントは AgentK とする．ベースラインには，先行研究 [1] の手法を使用する．

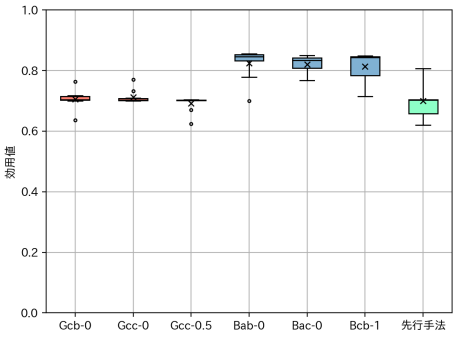

図 1: 提案手法と先行手法による獲得効用値の比較

図 1 は，学習に用いた相手エージェントごとの効用値平均が上位 3 つの場合と，先行手法を比較した結果である．この図では，相手に Gahboninho，報酬に式 (3-c)，式 (4-c)，−0.5 を使用したものは Gcc-0.5，相手に boulware，報酬に式 (3-a)，式 (4-b)，0 を使用したものは Bab-0 としている．本結果より，式 (3-a)，(3-c)，(4-b)，(4-c) は獲得効用が高くなる傾向があると言える．これは，交渉継続時に報酬を得ることでできるだけ交渉を続けて相手の譲歩を引き出し，さらに合意時の報酬が高くないので，より高い効用値を得ようと探索をするためである．また，boulware による学習は Gahboninho による学習よりも効用値が高くなる傾向にあり，図 1 に示した 3 つは先行手法と比較して，有意水準 1 %の Welch の t 検定により獲得効用が有意に高いことが確認された．これは，boulware の方が学習の際に多様な入力が与えられるためである．

5 まとめ

本研究では，自動交渉における受入戦略の深層強化学習フレームワークを提案した．実験により，先行手法よりも高い効用値が得られ，本手法の有効性が確認できた．

参考文献

[1] Yousef Razeghi, Celal Ozan Berk Yavuz, and Reyhan Aydoğan. Deep reinforcement learning for acceptance strategy in bilateral negotiations. *Turkish Journal of Electrical Engineering & Computer Sciences*, 28(4):1824–1840, 2020.

[2] Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 97–109, 1982.

[3] Shogo Kawaguchi, Katsuhide Fujita, and Takayuki Ito. Agentk: Compromising strategy based on estimated maximum utility for automated negotiating agents. In *New Trends in Agent-Based Complex Automated Negotiations*, pages 137–144. Springer, 2012.