キャラクターの役割分担のあるターン制 RPG の戦闘 AI
Turn-Based RPG Combat AI with Character Role Assignment

菊間　直樹
Naoki Kikuma
法政大学情報科学部コンピュータ科学科
E-mail:naoki.kikuma.8c@stu.hosei.ac.jp

Abstract

There are various studies in game AI including procedural content generation and non-player characters. These have been studied in many game genres including action and shooting games. However, there is less research on AI for role playing games (RPGs) than for other types of games. In this paper, we focus on the combat of a turn-based RPG and realize automatic combat using reinforcement learning. We propose a method that makes characters have individual roles and take tactical actions. In the RPG called Dragon Quest 4 (DQ4), characters on the player's side perform automatic combat based on multiple tactics and learn them in a rule-based manner. In contrast, the proposed method captures the importance of tactics in combat from the four perspectives of pinch, chance, the uniqueness of each role, and the compatibility of roles. Reinforcement learning using these four perspectives enables tactical automatic combat depending on the situation without relying on multiple strategies. In the experiment, we created a strategy "Gan Gan Ikouze" commonly used in DQ4, and compared the win rates of the agents using the proposed method against a boss character with those of "Gan Gan Ikouze" and the agents using random actions.

1. はじめに

近年のゲーム AI の進歩は著しい．例えばアクションゲームやシューティングゲームなどの Non-Player Character (NPC)に機械学習を取り入れることや，様々なアルゴリズムを用いてゲームコンテンツを生成する Procedural Content Generation (PCG)などが挙げられる．Nam ら [1]は PCG にアルゴリズムの 1 つとして機械学習を取り入れていた．このように多くのゲームジャンルで AI が研究されている．

様々なゲームジャンルの中で，Role Playing Game (RPG)がある．また RPG の中でも，アクション RPG やシミュレーション RPG など複数のジャンルがある．近年では RPG の研究は増え続けているが，強化学習を取り入れた研究は少なく，特にターン制 RPG の研究はあまり見られない．これは，アクション RPG や他のゲームジャンルと異なり，ターン制 RPG では強化学習を用いる場面が少ないからであると考える．アクション RPG は敵 NPC に強化学習を取り入れられるが，ターン制 RPG で敵 NPC は基本的にルールベースで行動したり，他の NPC に関しては戦う相手ではなく会話の対象であったりした．また従来のターン制 RPG はランダムで生成されるものが少ないため，強化学習を取り入れるのが困難であった．

本研究では，ターン制 RPG の戦闘で強化学習を用いてプレイヤー側のキャラクターが役割を分担して戦術的に戦闘を行う手法を提案する．RPG でキャラクターはそれぞれ役割に沿った固有のスキルを持つ．ここでは，その固有のスキルを持った役割のことをジョブと定義する．また異なるジョブで構成されたパーティが戦闘でどのような行動を取るのかを考えることを戦術と定義する．各ジョブが役割分担をすることで戦術的な戦闘を行うと考えられる．そのため役割分担の重要さをチャンス度，ピンチ度，ジョブのユニーク性，ジョブの相性の 4 つの要素から表現し，Q 学習による強化学習に導入する．実験では，ドラゴンクエスト 4 でよく使われる作戦である「ガンガンいこうぜ」を作成し，提案手法，ガンガンいこうぜ，ランダム行動を適用したエージェントでボスと戦闘させ，勝率とジョブの相性を比較する．

2. 関連研究

Nam ら [1]は PCG に機械学習を取り入れ，RPG でのダンジョンをバトルイベント，町などを非バトルイベントと捉え，それらを複数個まとめたものをステージとし，そのステージを多様かつ高品質に生成する手法を提案し，Deep Q-Network (DQN)や方策勾配法など様々な手法を用いて多様かつ高品質なステージを生成することに成功した．Kavanagh ら [2]はターン制 RPG でのゲームのバランス調整に連鎖戦略生成(CSG)と呼ばれる確率的モデル検査を提案した．これにより，ゲーム開発者が極端な戦略やプレイヤーが使いたいと思うかもしれないゲームプレイ方法を特定することに成功した．

3. 準備

3.1. Q 学習

Q 学習 [3]は強化学習の手法の 1 つである．基本的に強化学習は 6 つの要素がある．1 つ目はエージェントであり，学習を行う主体のことである．2 つ目は環境であり，エージェントが学習するための場所である．3 つ目は状態であり，エージェントの状態を示している．4 つ目は行動であり，エージェントが選択できる行動を示している．5 つ目は報酬であり，行動に対する報酬を設定するものである．6 つ目は方策であり，状態sである行動aをとる確率のことである．ここで，ある状態である行動を取った時の価値を状態行動価値といい Q 値と呼ばれる．Q 学習は Q 値を更新させて最適な行動を学習する．

Supervisor: Prof. Hiroshi Hosobe

3.2. ターン制 RPG の戦闘

ターン制 RPG の戦闘をプレイヤー側に有利に運ばせる方法は 2 つあると考えられる．1 つ目は戦略性，2 つ目は戦術性である．戦略性は，戦闘準備段階でどのようなパーティで戦闘に挑むのか，またどのような装備をしていくのかなどを考えることである．戦術性は，戦闘で自分のターンが回った時にどのような行動をとるのかを考えることである．目的は戦闘で状況に応じた行動を強化学習で学習させることであるため，戦術的行動を取ることを目指しているといえる．本研究では装備を考慮していないが，パーティの構成を考慮しているため，戦略的とも捉えられる．そのため本研究では，あるパーティ構成で状況に応じた行動を取ることを戦術と定義する．

3.3. 自動戦闘における作戦命令

本研究で作成する自動戦闘はドラゴンクエスト 4 (DQ4) の作戦命令を想定している．種類は様々あり，例として「ガンガンいこうぜ」という作戦を選択すると，味方は強力なスキルを使うように動く．また「いのちを大事に」を選択すると，味方は防御や回復スキルを使うように動く．DQ4 にはこの作戦命令で戦闘すると，戦闘ごとに味方が学習する．この学習はルールベース[1]で実装されている．

4. 提案手法

本研究では Q 学習を用いてキャラクターが役割分担をして戦術的に戦闘を行う手法を提案する．

4.1. 概要

キャラクターの役割をジョブと呼び，ジョブの重要さを以下の 4 つの要素で表現し導入する．1 つ目はピンチ度である．ピンチ度は現時点でのプレイヤー全体の HP や MP から計算したものである．ピンチ度に応じて，プレイヤーの選択する行動が変化すると考える．以下がピンチ度の計算式である．total_hp_mp は味方全体の現在の HP と MP の合計である．max_hp_mp は味方全体の最大 HP と MP の合計である．

$$\text{pinch} = 1 - \text{total\_hp\_mp}/\text{max\_hp\_mp}$$

2 つ目はチャンス度である．チャンス度は現時点で敵に与えたダメージの合計と経過ターン数から計算したものである．敵の HP の正確な情報は得られないが，チャンス度に応じてもう少しで倒せるかもしれないという状況でプレイヤーの行動に変化をもたらすと考える．以下はチャンス度の計算式である．ave_damage は 1 ターンの平均ダメージである，max_hp_enemies は敵の最大 HP である．

$$\text{chance} = \text{ave\_damage}/\text{max\_hp\_enemies}$$

3 つ目はジョブのユニーク性である．それぞれのジョブは固有のスキルをもち，ステータスもジョブによって異なる．それらを考慮してジョブを選択することで戦術的な戦闘を行えると考える．そのためユニーク性は各ジョブのステータスやスキルで表現し，それを報酬の計算に用いることでそれぞれのジョブの行動が戦術的に近づくと考える．

4 つ目はジョブの相性である．ユニーク性に関連しており，ユニーク性は個々のキャラクターに対してのものだが，相性はプレイヤー側のチーム全体に焦点を置いている．これを考慮したチーム選択はジョブ同士のシナジー効果を生み出し，さらに戦術性を伴った戦闘が行えると考える．そのため，ジョブの相性は報酬の合計値で表す．各ジョブのステータスやスキルを加味した報酬の計算をしているため，その合計値が高い値であれば戦術的であるといえる．つまり役割分担を可能としていると考える．

ピンチ度とチャンス度を 10 段階ずつに分け，それを状態とする．以下がその計算式である．state が状態を表す．

$$\text{pinch\_dis} = \min(\text{floor}(\text{pinch} \times 10), 9)$$
$$\text{chance\_dis} = \min(\text{floor}(\text{chance} \times 10), 9)$$
$$\text{state} = (\text{pinch\_dis}, \text{chance\_dis})$$

ピンチ度とチャンス度は報酬の計算にも用いる．報酬に関して詳しくは次節で記述する．行動はプレイヤーが選択できるものとし，通常攻撃とスキルを選択できる．

4.2. 報酬の計算

報酬では，ジョブのユニーク性を活かしてスキルの使用や回復，バフの使用する対象によって報酬が変わるようにする．基本的には勝利すると多くの報酬がもらえ，プレイヤーが全滅すると報酬が減るようにする．しかし，相手にどのくらいダメージを与えられているのかによっても報酬がもらえるようにする．これらの報酬の合計がジョブの相性となり，戦闘の全体の流れで戦術的な行動がとれているかがわかる．

報酬の具体的な値は以下の計算式で求める．reward は報酬の合計値である．damage はプレイヤーが敵に与えるダメージである．enemy_max_hp は敵の最大 HP である．heal_amount は回復量を表しており，回復するプレイヤーの最大 HP の 4 割である．target_max_hp は回復するプレイヤーの最大 HP を表す．

- 通常攻撃

  $\text{reward} = \text{reward} + \text{damage}/\text{enemy\_max\_hp} \times 50$
- スキル使用
  1. MP 不足の場合

     $\text{reward} = \text{reward} - 20$
  2. 回復スキル

     $\text{reward} = \text{reward} + \text{heal\_amount}/\text{target\_max\_hp} \times 30$
  3. HP が 5 割以下の時に回復

     $\text{reward} = \text{reward} + 100$
  4. バフをかける

     $\text{reward} = \text{reward} + 50$
  5. 攻撃スキル

     $\text{reward} = \text{reward} + \text{damage}/\text{enemy\_max\_hp} \times 50$
  6. バフ中に攻撃スキルを使用

     $\text{reward} = \text{reward} + 50$
- 状況に応じた報酬

[1] https://crimson-ytb.hatenablog.com/entry/2022/04/02/183328

1. ピンチ度 0.7 以上の時に回復スキルを用いてピンチを打破

$$\text{reward} = \text{reward} + \text{pinch} \times 30$$

2. チャンス度 0.7 以上の時に攻撃することでチャンスを生かした行動を選択

$$\text{reward} = \text{reward} + \text{chance} \times 30$$

- 勝敗による報酬
  1. 勝ち

$$\text{reward} = \text{reward} + 100$$

  2. 負け

$$\text{reward} = \text{reward} - 100$$

5. 実装

強化学習を取り入れるためのターン制 RPG の戦闘システムを作成した．作成した戦闘システムは，DQ4 の戦闘を模範としている．プレイヤーと敵が攻撃やスキルを駆使して，HP が 0 になるまで戦闘を行う．プレイヤー側は 3 人で，それぞれ固有のスキルを持ったジョブを持っている．敵側は 1 体であり，ボス的存在であるため，強力なスキルやステータスを持っている．各ジョブとボスのステータスを表 1 に，スキルとスキルの効果を表 2 と表 3 に示す．スキルに関してプレイヤー側は 2 つ持っているのに対して，ボスは 3 つ持っている．またボスはランダム行動ではなく簡単なルールベースで動くようにした．以下がボスの行動である．

- HP100%～50%の時
  1. 通常攻撃とスキル「スラッシュ」を同確率で行う．
- HP50%未満の時
  1. 最初にスキル「バフ」をかける(1 度だけ)
  2. 通常攻撃とスラッシュを同確率で行う．
  3. 通常攻撃やスラッシュの 2 分の 1 の確率でスキル「クリーブ」を使用する．
- MP 不足の時
  1. 通常攻撃を行う．

実装した Q 学習はプレイヤー側全体をエージェントとしている．状況はピンチ度とチャンス度で表現している．行動は，各プレイヤーは 2 つのスキルを所持しているため，攻撃と各スキルの 3 つの行動を選択できる．方策にはε-greedy 法を用いる．これは 0 から 1 の範囲にあるεを決めておき，εの確率でランダム行動をし，それ以外で Q 値に基づいて最適な行動をする方策である．本研究では，これを少し応用し学習が進むにつれてεを減少するようにした．つまり，学習初期段階はランダム行動しやすく，学習が進むと Q 値が高いものが選択されやすくなっている．Q 学習の学習率は 0.1 で，割引率は 0.99 としている．

比較実験に使用する「ガンガンいこうぜ」を作成した．ガンガンいこうぜは，基本的にはその時点で使用可能な最も強いスキルを使用し続けるものである．DQ4 のガンガンいこうぜでは回復スキルを使わない場合もあるが，作成したものでは現在 HP が最大 HP の 10%を下回れば回復を行うようにした．

表 1. プレイヤーと敵のステータス

名前	HP	MP	攻撃力	魔力	物理防御力	魔法防御力
ボス	900	100	37.5	30.0	14.4	12.0
戦士	300	50	30.0	10.0	14.0	6.0
僧侶	180	100	16.0	24.0	10.0	15.0
魔法使い	150	120	16.0	30.0	8.0	12.0

表 2. スキル

名前	スキル 1	スキル 2	スキル 3
ボス	スラッシュ	バフ	クリーブ
戦士	スラッシュ	サンダースラッシュ	
僧侶	ヒール	バフ	
魔法使い	ファイアボール	ライトニング	

表 3. スキルの効果

スキル名	効果
スラッシュ	単体物理攻撃，攻撃力の 1.3 倍，MP 消費 5
サンダースラッシュ	単体物理攻撃，攻撃力の 1.8 倍，MP 消費 10
ヒール	プレイヤー1 人の最大 HP の 4 割回復する．MP 消費 8
バフ	プレイヤー1 人の攻撃力を 2 倍にする．MP 消費 8
ファイアボール	単体魔法攻撃，魔力の 1.4 倍，MP 消費 5
ライトニング	単体魔法攻撃，魔力の 2.0 倍，MP 消費 10
クリーブ	全体物理攻撃，攻撃力の 1.3 倍，MP 消費 8

6. 実験

作成した戦闘 AI が役割分担を可能とする戦闘をおこなうかどうかを検証するために 2 つの実験を行った．比較実験では，ガンガンいこうぜを積極的な行動をとるエージェントとして捉え，図では Aggressive Agent と表記している．

6.1. 実験 1

実験 1 ではガンガンいこうぜとランダム行動，作成したエージェントがボスと戦闘し，その勝率の比較実験を行う．また，ガンガンいこうぜと Q 学習の報酬の遷移図も比較する．本研究では報酬の合計値がジョブの相性や戦術性を表現しているため，ガンガンいこうぜと学習後を比較することで，既存の作戦命令より戦術的であり，役割分担を可能としているといえる．

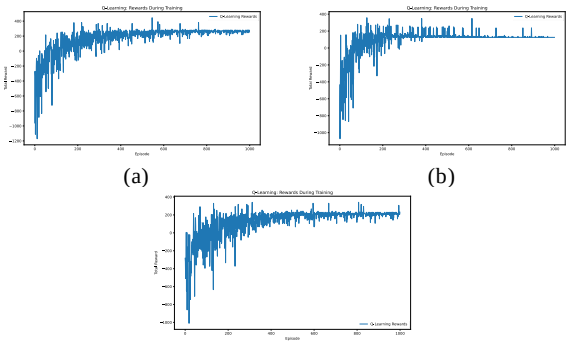
6.2. 実験 1 の結果

1000 回の戦闘で学習を行った．図 1 は報酬の学習遷移図である．報酬の合計値はジョブの相性であるため，学習するにつれて戦術的な行動をとれていることがわかる．(a)～(c)はそれぞれ 1～3 回目の学習遷移図である．次に勝率の比較実験を行ったところ，ボスとの戦闘でランダム行動はすべて 0%であった．このため，ランダム行動を除く勝率を以下の表 4 に示す．敵や味方のパラメータを様々変更したが，大きな変化はなかった．しかし，戦闘として成立している．敵と味方のパラメータやスキルの

バランスについて深く追求することは本研究の目的ではないため，以降は学習後のエージェントとガンガンいこうぜの報酬の遷移のみで戦術的かどうかを比較する実験を行った．図 2 はガンガンいこうぜと学習後の報酬の遷移図である．戦闘回数は 500 回である．また表 4 はその時のそれぞれの勝率を表している．勝率が低い時，ジョブの相性が機能していないが，勝率が 100%の時は，ガンガンいこうぜより相性が良いことがわかる．1 回目と 3 回目は勝率が低くまた報酬の合計値も低いことから学習がうまくいっていない可能性があると考える．同様の実験を何度か行ったが，1 回目と 3 回目のような結果と 2 回目のような結果のどちらも得られた．

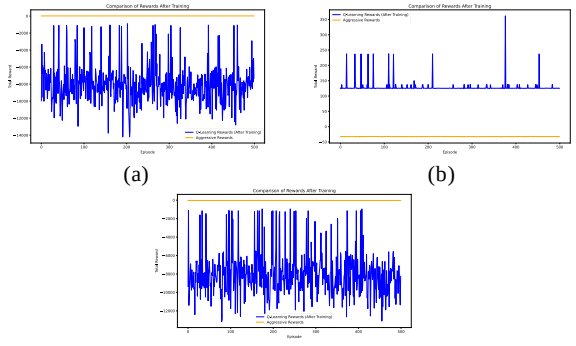
表 4. ガンガンいこうぜと学習後のエージェントの勝率 (戦闘回数 500 回)

	1 回目	2 回目	3 回目
提案手法	13.2%	100.0%	11.6%
ガンガンいこうぜ	100.0%	100.0%	100.0%



(a) (b) (c)

図 1. 報酬の遷移：(a) 1，(b) 2，(c) 3 回目



(a) (b) (c)

図 2. 学習後のエージェントとガンガンいこうぜのジョブの相性の比較：(a) 1，(b) 2，(c) 3 回目

6.3. 実験 2

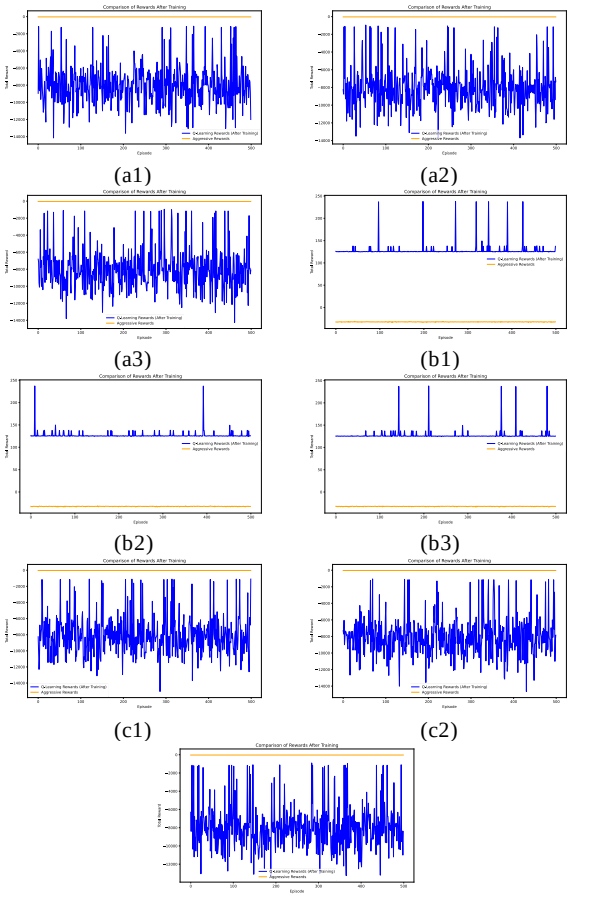
実験 2 では，実験 1 で行った学習結果を用いて，学習後の戦闘を何度か行い，その時の勝率と報酬の合計値の比較を行う．これにより，学習がうまくいっているのか，また学習後のエージェントの挙動に問題があるかなどを確認できると考える．ここではガンガンいこうぜとの比較ではなく，学習できているのかを比較するため，図の青い線に注目する．

6.4. 実験 2 の結果

図 3 (a1)～(a3), (b1)～(b3), (c1)～(c3)は図 1(a), (b), (c)にそれぞれ対応している．つまり図 3(a1)～(a3)が実験 1 で 1 回目に学習したもので，(a1)はその学習結果を用いた戦闘を行った結果を表している．(b1)～(b3)と(c1)～(c3)も同様である．図 3 から，(a1)～(a3)と(c1)～(c3)は報酬を低く出力し，かなりばらつきがあることがわかる．一方，(b1)～(b3)は高い値を出力していて，期待通りの結果が得られた．表 5 はそれぞれの勝率を表している．図 3，表 5 から学習後の戦闘の挙動に問題がある可能性は低く，学習中の挙動に差異が多く生じている可能性が高いと考える．よって，実験 1 の 1 回目と 3 回目はうまく学習が行えていなくて，2 回目はうまく学習が行えていることがわかった．

表 5. 提案手法の勝率

実験 1 との対応	1 回目	2 回目	3 回目
(a)	10.8%	12.4%	12.0%
(b)	100.0%	100.0%	100.0%
(c)	11.2%	13.0%	13.2%



(a1) (a2) (a3) (b1) (b2) (b3) (c1) (c2) (c3)

図 3. 報酬の合計値の遷移

6.5. 実験 3

実験 3 ではパーティ構成を少し変えた状態で，実験 1 と同様の比較実験を行う．本研究では，簡略化のためジョブの数を最低限にしている．そのため，ジョブの種類は 3 種類のジョブで構成されるいくつかの異なるパーティ構成で実験を行う．3 種類のジョブしかないため，パーティの差異により大きな変化が得られるかはわからないが，これによりパーティごとに変わった役割分担をする必要がでるため，パーティごとに報酬の差異が生じると考える．

6.6. 実験 3 の結果

パーティの構成は以下の表 6 に示す．図 4 はそれぞれのパーティの学習後の戦闘での報酬の合計値である．(a)がパターン 1，(b1)～(b3)がパターン 2，(c1)～(c3)がパターン 3 に対応している．学習に関して，実験 1 と 2 を踏まえうまく学習が行えていると考えるものを選択している．実験 1 や 2 で扱ったパーティ構成よりはバランスが劣るが，ある程度バランスが良いとされている表 6 のパターン 2 とパターン 3 の構成に対して実験を 3 回行った．パターン 1 は非常に攻撃的なパーティ構成ではあるが，役割分担を考える必要性が他の構成に比べて少ないため，他のパターンより実験の試行回数を少なくしている．勝率に関しては，パターン 1 とパターン 2 の構成はすべて 0%で，パターン 3 の構成は 100%であった．そのため，図 4 よりパターン 3 の構成ではガンガンいこうぜより報酬がよい場合もあった．

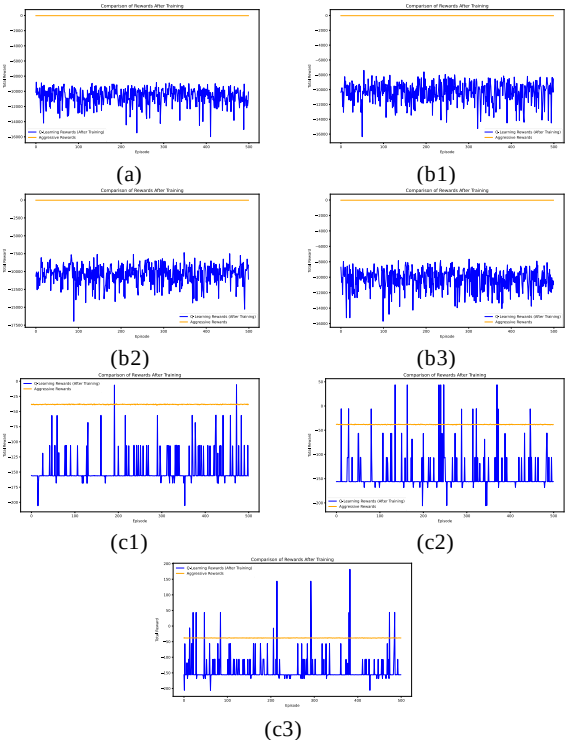


図 4. 異なるパーティ構成での報酬の合計値の遷移

表 6. パーティ構成

	パーティ構成		
パターン 1	戦士	戦士	戦士
パターン 2	戦士	戦士	僧侶
パターン 3	魔法使い	魔法使い	僧侶

7. 議論

実験 1 から，ガンガンいこうぜより学習後のエージェントの方が良い結果が得られた．しかし，期待通りの結果が得られたとは言い難い．図 1 から，1～3 回目はどれも 200 エピソード前後で大きく負の値を出力している．これは，ほぼランダムに近い行動を取っていて，MP 不足によるスキル使用が多発していると考えられる．また 400 エピソード超えたあたりから報酬の値が安定している．しかし，安定していても増えたり減ったりしているエピソードがあるのは，ダメージ計算に乱数を用いているからだと考えられる．図 2 (b)からは，学習後のエージェントがガンガンいこうぜよりも報酬が多く得られていることがわかる．また表 4 によると勝率が 100%である．それに対して，図 2(a), (b)からは，ガンガンいこうぜより低いことがわかる．また表 4 から，1 回目と 3 回目は勝率が低いことがわかる．このことから勝率とジョブの相性が比例していることがわかる．図 1 から，多少の差異はあるが，この時点でうまく学習できていない可能性は低いと考える．1 回目と 3 回目の報酬が低かったり，勝率が低かったりするのは学習後に行った戦闘に問題があると考える．しかし，2 回目は勝率も報酬も期待通りに出力している．バフ中の攻撃スキルの使用や回復のタイミグによる追加報酬があるが，1 回目と 3 回目で負の値が極端に出力されていることから，MP 不足に原因があると考える．また図 2 からガンガンいこうぜの得られた合計値が極端に少なく報酬が 0 を下回っている．勝利による報酬は 100 もらえるが，それより下回っているということは，負の報酬が多かったといえる．2 つのエージェントの比較から，MP 不足によるスキル使用によって多くの戦闘で報酬が減っているのではないかと考える．

実験 2 からは実験 1 で考えられる可能性とは違う可能性が考えられることがわかった．実験 1 では学習後に問題があると考えた．しかし，実験 2 からは学習後の挙動に問題がないことがわかった．したがって，問題があるのは学習中である可能性が高いと考える．実験 1 と同様に図 3 から負の報酬が顕著に表れているため，負の計算に問題があると考える．

実験 3 から，パターン 3 の構成のみガンガンいこうぜより良い値を出力したエピソードがあることを確認できた．他のパターンに関して，負の値を頻繁に出力しているため学習時点でうまくいっていないのではないかと考え，何度か学習を行ったうえで，学習後の戦闘実験を行っている．そのため，パーティ構成が悪くガンガンいこうぜより悪い値を出力している可能性が高いと考える．その他の要因としても，表 6 のパターンは全体的に攻撃的なパーティ構成であるので，積極的に強力なスキルを使うガンガン行こうぜの方が優れているのではないかと考える．

実験 1〜3 を通して，負の報酬の計算に問題が発生している可能性が高いことがわかった．本研究に用いた戦闘システムで頻繁に発生する負の報酬の計算は MP 不足によるスキル使用である．そのため，MP 不足によるスキル使用の点でうまく学習できていないのではないかと考える．主な負の報酬の計算が MP 不足によるスキル使用しかないので，改善案として他に負の報酬を得る要素を増やすことで MP 不足によるスキル多用を現時点より緩和できると考える．その他の点でも，報酬や戦闘システムの見直しをすることで，現状よりもうまく学習できていない回数が減ると考える．

8. おわりに

本研究ではターン制 RPG で Q 学習を用いて役割分担を可能にする戦闘 AI を提案した．提案した手法を用いて実装したエージェントは，うまく学習が行えている場合，作成した環境に適応したガンガンいこうぜよりは戦術的であるという結果が得られた．作成した環境は DQ やファイナルファンタジー(FF)など一般的なターン制 RPG に比べて簡略化されたものである．ステータスのパラメータの数，スキルの数やジョブの数など多くの点でより複雑な戦闘システムにする要素がある．

今後の課題は，本研究の問題を解決したうえでより複雑な戦闘システムで役割分担を可能とした戦闘 AI を作成することである．しかし，研究結果からプレイヤー全体を Q 学習で扱うことは難しいと考えられる．本研究の結果が得られたのは簡略化された戦闘システムであったからだと考える．そのため，より複雑な戦闘システムで，プレイヤー全体を扱うのであれば DQN や深層方策勾配法などの深層学習を用いて複雑な内容でも扱えるような強化学習を用いるべきだと考える．また，プレイヤー全体を学習させるのではなく，各プレイヤーが個々に学習していければさらに発展できると考える．これをマルチエージェントといい，Cruz ら [4]はボードゲームでマルチエージェントを用いていた．ターン制 RPG の戦闘にマルチエージェントを用いることで複雑なシステム，つまり DQ や FF など近年のターン制 RPG にも組み込まれている戦闘システムでも役割分担を実現させられるのではないかと考えられる．

文　献

[1] S. G. Nam, C. H. Hsueh and K. Ikeda, "Generation of Game Stages With Quality and Diversity by Reinforcement Learning in Turn-Based RPG," *IEEE Transactions on Games*, vol. 14, no. 3, 2022.

[2] W. Kavanagh, A. Miller, G. Norman and O. Andrei, "Balancing Turn-Based Games With Chained Strategy Generation," *IEEE Transactions on Games*, vol. 13, no. 2, 2021.

[3] C. J. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3-4, pp. 279-292, 1992.

[4] D. Cruz, J. A. Cruz and H. L. Cardoso, "Reinforcement Learning in Multi-agent Games:Open AI Gym Diplomacy Environment," *EPIA*, vol. 1, pp. 49-60, 2019.