不誠実な論証を扱う対話モデルの評価及び説得戦略の考察

Evaluation of a Dialogue Model for Untrusted Arguments and Discussion on a Strategy for Success of Persuasion

國生　一幸[1*]　高橋　和子[1]
Kokusho Kazuyuki[1]　Takahashi Kazuko[1]

[1] 関西学院大学大学院理工学研究科
[1] Graduate School of Science and Technology, Kwansei Gakuin University

Abstract: This work aims at dealing with persuasive dialogues including the situation in that an agent intentionally hides a part of her knowledge or in that her opponent reveals it. We consider a dialogue model in which each agent has her own knowledge base and prediction of her opponent's knowledge base represented in the form of an argumentation framework. We define a dialogue protocol allowing dishonest arguments, implement the model and evaluate it. As a result, we have clarified that the result of persuasion depends on the agent's initial knowledge. We have also found that dishonest arguments and prediction of the opponent's knowledge can increase the number of dialogues in which persuasion succeeds or the opponent's dishonesty is revealed. In addition, we obtain a new strategy from the analysis of this experimental result. We can increase more the number of dialogues in which the ratio of dialogues advantageous to the persuader.

1　はじめに

現実の対話で起こっていることを理解するために、対話モデルの研究は人工知能を中心とした様々な分野でさかんに行われている。中でも議論は対話の 1 つと見なされ、論理的要素を多く含む。対話モデルはこれまで多く提案されているが、評価実験についてはまだ十分なされているとはいえない。本研究で扱う議論とは議題に関しての発言とその発言に対する反論を繰り返すことによって、構築される一連の対話である。Dung は議論の発言とその反論をグラフにおけるノードとエッジに対応させることにより、議論の構造を抽象化する議論フレームワーク (AF) を提案した [3]。AF は発言である論証と発言に対する反論である攻撃関係から構成される。Amgoud らは AF を使って、各エージェントが独立した知識ベースをもち、対話の進行とともにそれが更新される対話モデルを提案した [1]。Sakama は対話の過程で人間が不正直な発言をする心理状態をモデル化するために、不正直な発言を含む論争ゲームを考案した [4]。Yokohama は Amgoud らのモデルに加えて、相手の知識を予測する知識を与えることにより、不正直な発言を指摘する手法や相手に反論を与えないようにする戦略を提案した [5, 6]。ここでいう戦略とは、複数の可能な発言から 1 つの発言を選択することである。

たとえば、学生が研究室を選択する例を考える。この例では不正直な発言を指摘するときに、予測知識をどのように用いるのかを示している。アリスはボブと同じ研究室に入りたいと考えている。彼女はチャーリー教授が優しいが、面倒見がよくない教授であることを知っている。一方ボブは、厳しいか、または面倒見がよい教授の研究室に入りたいと考えている。まず、ボブがチャーリー教授について何も知らない場合について考える。アリスが「一緒にチャーリー研に入ろう」と発言し、ボブは「チャーリー教授がどのような人物か分からないままでは、一緒に入ることはできない」と発言したとする。このときに、アリスが「チャーリー教授は優しいから一緒に入ろう」と発言したとする。ボブは「僕は厳しい教授がいいんだ」と発言し、アリスはその後反論できないため、説得に失敗する。では、アリスが「チャーリー教授は面倒見がよい」と発言したとするとどうなるだろうか。ボブはアリスから「チャーリー教授は面倒見がよい」ということを聞き、ボブは面倒見のよい教授の研究室に入りたいため、アリスは説得に成功する。アリスの説得が成功したのは「チャーリー教授は面倒見がよくない」という事実を隠して、「チャーリー教授は面倒見がよい」という不正直な発言をしたからである。次に、ボブが「アリスは『チャーリー教授は面倒見がよくない』と知っている」ことを知っていた

*連絡先：関西学院大学大学院 理工学研究科
〒 669-1337 兵庫県三田市学園 2 丁目 1 番地
E-mail: dbq68680@kwansei.ac.jp

とする。アリスが「チャーリー教授は面倒見がよい」と発言したとする。このとき、ボブは「アリスは『チャーリー教授は面倒見がよくない』と知っている」ことを知っているため、「『チャーリー教授は面倒見がよくない』と知っていたのにもかかわらず、『チャーリー教授は面倒見がよい』と発言したんじゃないのか」と不正直な発言を指摘することができる。このように、相手の知識を予測する知識を与えることにより、相手の不正直な発言を指摘することが可能となる。また、説得の成功はエージェントが知っている知識と選択する発言によって決まる。

Yokohama はこのような場合に、相手を説得するための戦略を提案した。しかし、提案した対話モデルの評価実験をしていないため、不正直な発言をすることや戦略を用いることに優位性があるのかを明確にしていない。そこで、本研究では Yokohama の提案した対話モデルの評価実験をすることで、エージェントが持つ知識や選択した発言が対話にどのような影響を与えているのかを調査する。

本発表の構成は以下の通りである。第 2 節では対話モデルで用いる議論システムの定義について述べる。第 3 節では評価実験に用いる対話モデルについて述べる。第 4 節では対話モデルの評価実験について述べる。第 5 節では本研究の成果と今後の課題について述べる。

2 議論システム

本研究で使用する、AF は $AF{=}(AR, AT)$ と定義される [3]。AR は論証の集合、$AT \subseteq AR \times AR$ は攻撃関係である。ラベリング L は、AR から $\{in, out, undec\}$ への関数である [2]。ラベリングが任意の $A \in AR$ に対して、下記の条件を満たすとき、完全ラベリングという。

- $L_{AF}(A) = in \Leftrightarrow$
  $(\forall B \in AR;\ (B, A) \in AT \Rightarrow L_{AF}(B) = out)$
- $L_{AF}(A) = out \Leftrightarrow$
  $(\exists B \in AR;\ (B, A) \in AT \wedge L_{AF}(B) = in)$
- $L_{AF}(A) = undec \Leftrightarrow (L_{AF}(A) \neq in \wedge L_{AF}(A) \neq out)$

AF に対する完全ラベリングにおいて、ラベリングが in である論証の集合を $in(L_{AF})$ と表記する。

$$in(L_{AF}) = \{A | L_{AF}(A) = in\}$$

$in(L_{AF})$ が極小であるようなラベリングを基礎ラベリングという。任意の AF に対して、基礎ラベリングが 1 つだけ存在する [2]。

3 対話モデル

3.1 議論システムに基づく対話モデル

Yokohama は Amgoud が提案した対話モデルに相手の知識を予測する知識ベースの概念を加えることにより新しい対話モデルを提案した [6]。現実に即した対話モデルを構築するために、Yokohama の提案した対話モデルは正直な発言だけでなく不正直な発言も扱っている。各エージェントは相手の知識を予測する予測知識ベースも所持している。予測知識ベースを用いることにより、不正直な発言を指摘することが可能となる。

$AF{=}(AR, AT)$、$AF_S{=}(AR_S, AT_S)$ とする。
$AR_S \subseteq AR$ かつ $AT_S = AT \cap (AR_S \times AR_S)$ であるならば、AF_S を AF の部分 AF と定義し、$AF_S \subseteq AF$ と記述する。対話モデルにおけるすべての知識を表す AF を UAF とよぶ。

エージェント $\mathbf{X}$、$\mathbf{Y}$ の AF をそれぞれ $AF_X = (AR_X, AT_X)$、$AF_Y = (AR_Y, AT_Y)$ とするとき、AF_X、AF_Y は UAF の部分 AF である。$\mathbf{Y}$ の知識を予測する $\mathbf{X}$ の知識ベース PAF_Y は $PAF_Y \subseteq AF_X$、$PAF_Y \subseteq AF_Y$ を満たす。相手の知識を予測する場合、自分もその知識を持っている必要があるため、$PAF_Y \subseteq AF_X$ という条件を付加する。また、予測する知識が間違っていないと仮定しているため、$PAF_Y \subseteq AF_Y$ という条件を付加する。

Act はエージェントの発言の種類を表す。$Assert(\rho, -)$, $Assert(A, B)$ は言明であり、議題 ρ の発言、または相手エージェントが過去に発言した意見 B を攻撃する発言 A である。$Suspect(A, B)$ は相手エージェントが過去に発言した、不正直な発言を指摘する。$Excuse(A, B)$ は相手エージェントの不正直な発言の指摘に対する弁解である。$Pass$ は何も発言しないことである。Act を T としたときに、手を (X, T) と定義する。Act から論証の集合への関数 F_{cs} は $Assert(\rho, -)$ なら ρ、$Assert(A, B)$、$Suspect(A, B)$、$Excuse(A, B)$ なら A、$Pass$ なら $\emptyset$ を返す。

$\mathbf{P}$ を説得者、$\mathbf{C}$ を被説得者、ρ を説得者 $\mathbf{P}$ が最初に発言する論証とする。ρ に関する $\mathbf{P}$ と $\mathbf{C}$ の対話を $d_0 = [\,]$、$d_k = [m_0, ..., m_{k-1}] (1 \leq k)$ と定義する。対話における各手 $m_i (0 \leq i \leq k-1)$ は (X_i, T_i) の形を成し、以下の条件を満たす。

- $T_0 = Assert(\rho, -)$
- i が偶数のとき $X_i = P$、i が奇数のとき $X_i = C$
- m_i は可能手である。(後ほど定義する。)

いずれかの条件を満たすと対話が終了する。

- Pass による対話終了 $(3 \leq k)$
  $m_{k-2} = (X, Pass)$ かつ $m_{k-1} = (Y, Pass)$
- Suspect による対話終了 $(3 \leq k)$
  $m_{k-2} = (X, Suspect(A, B))$ かつ $m_{k-1} = (Y, Pass)$

d_k における $\mathbf{X}$ のコミットメントストア $CS_X^{d_k}$ を $k = 0$ ならば $CS_X^{d_k} = \emptyset$、$k \neq 0$ ならば $CS_X^{d_k} = \bigcup_{i=0,...,k-1, X_i=X} F_{cs}(T_i)$ と定義する。$CS_X^{d_k}$ では、$\mathbf{X}$ がこれまでに発言した論証を知識として保持している。

対話 d_k で $\mathbf{X}$ が所持する議論フレームワーク $AF_X^{d_k} = (AR_X^{d_k}, AT_X^{d_k})$ を $AR_X^{d_k} = AR_X \cup CS_Y^{d_k}$, $AT_X^{d_k} = AT \cap (AR_X^{d_k} \times AR_X^{d_k})$ と定義する。また、対話 d_k で

$\mathbf{X}$ が所持する $\mathbf{Y}$ の予測議論フレームワーク $PAF_Y^{d_k} = (PAR_Y^{d_k}, PAT_Y^{d_k})$ を $PAR_Y^{d_k} = PAR_Y \cup CS_X^{d_k} \cup CS_Y^{d_k}$、$PAT_Y^{d_k} = AT \cap (PAR_Y^{d_k} \times PAR_Y^{d_k})$ と定義する。$AF_X^{d_k}$ は $\mathbf{X}$ が対話 d_0 時点で所持している AF_X と $\mathbf{Y}$ が対話 d_k 時点までに発言した論証により構成されている。$PAF_Y^{d_k}$ は $\mathbf{X}$ が対話 d_0 時点で所持している PAF_Y と $\mathbf{X}$ と $\mathbf{Y}$ が対話 d_k 時点までに発言した論証により構成されている。

$A \in AR_X^{d_k}$、$B \in CS_Y^{d_k}$ とする。$m_k = (X, Assert(A,-)/Assert(A,B)/Excuse(A,B))$ であるとき

- $L_{AF_X^{d_k}}(A) = in$ ならば m_k は正直な発言である。
- $L_{AF_X^{d_k}}(A) \neq in$ ならば m_k は不正直な発言である。

ここで不正直な発言とは、自分に都合の悪いことは意図的に発言しないという隠蔽を意味する。

d_k における $\mathbf{X}$ の各手 $m_i(0 \leq i \leq k-1)$ が以下の条件を満たすとき、m_i を $\mathbf{X}$ の可能手とする。

手に共通する定義を以下に示す。

- パスが二回続いた場合は対話が終了する。
- 対話の中でパス以外に同じ手は存在しない

上記に加え、各手を出せる条件は以下の通りである。

- 言明
  - $k = 0$ ならば $(P, Assert(\rho, -))$
  - $k \neq 0$、$m_{k-1} \neq (Y, Suspect(C,D))$、$A \in AR_X^{d_k}$、$B \in CS_Y^{d_k}$、$(A,B) \in AT_X^{dk}$ ならば、$(X, Assert(A,B))$
- 不正直な発言の指摘
  - $m_{k-1} = (Y, Assert(\rho,-))$、$\rho, B \in PAR_Y^{d_k}$、$(B,\rho) \in PAT_Y^{d_k}$、$L_{PAF_Y^{d_k}}(\rho) \neq in$ ならば、$(X, Suspect(B,\rho))$
  - $m_{k-1} = (Y, Assert(B,A))/(Y, Excuse(B,A))$、$B, C \in PAR_Y^{d_k}$、$(C,B) \in PAT_Y^{d_k}$、$L_{PAF_Y^{d_k}}(B) \neq in$ ならば、$(X, Suspect(C,B))$
- 弁解
  - $k \neq 0$、$m_{k-1} = (Y, Suspect(B,A))$、$C \in AR_X^{d_k}$、$B \in CS_Y^{d_k}$、$(C,B) \in AT_X^{dk}$ ならば、$(X, Excuse(C,B))$
- パス
  - $k \neq 0$ ならば、$(X, Pass)$ 何も発言しない

手 $m_k = (X,T)$ を出した後、以下の順で更新がなされる。

1. 対話の更新
   $d_{k+1} = [m_0, ..., m_{k-1}, m_k]$
2. コミットメントストアの更新
   $CS_X^{d_{k+1}} = CS_X^{d_k} \cup F_{cs}(T)$
3. $\mathbf{X}$ が所持している AF の更新
   $PAF_Y^{d_{k+1}} = (PAR_Y^{d_k} \cup CS_X^{d_{k+1}}, AT \cap (PAR_Y^{d_{k+1}} \times PAR_Y^{d_{k+1}}))$
4. $\mathbf{Y}$ が所持している AF の更新
   $AF_Y^{d_{k+1}} = (AR_Y^{d_k} \cup CS_X^{d_{k+1}}, AT \cap (AR_Y^{d_{k+1}} \times AR_Y^{d_{k+1}}))$
   $PAF_X^{d_{k+1}} = (PAR_X^{d_k} \cup CS_X^{d_{k+1}}, AT \cap (PAR_X^{d_{k+1}} \times PAR_X^{d_{k+1}}))$

下記の条件を満たすとき、勝敗が決定する。

1. *Pass* により対話が終了したとき
   (a) $L_{AF_C^{d_k}}(\rho) = in$ であるならば、説得が成功したことにより $\mathbf{P}$ が勝利する。
   (b) $L_{AF_C^{d_k}}(\rho) \neq in$ であるならば、説得が失敗したことにより $\mathbf{P}$ が敗北する。
2. *Suspect* により対話が終了したとき
   相手の不正直な発言を見破ったことにより $\mathbf{X}$ が勝利する。

上記の 3 つの条件は同時に満たされることはない。

3.2 戦略

対話 d_k においてエージェントの手 m_k の選び方の優先順位を変えた戦略を 3 つ考える。

戦略 S_Y は Yokohama が提案した戦略 [6] を改良したものである。戦略 S_Y では、$\mathbf{P}$ が対話で敗北してしまうような可能手を避け、相手が議題を信じているとわかったら、相手に情報を開示しないことで自分が不利になる可能性を減らそうとする。

戦略 $\mathbf{S_Y}$

1. $m_{k-1} = (C, Suspect(B,A))$ であるとき、$\rho \in PAR_C^{d_{k+1}}$、$L_{PAF_C^{d_{k+1}}}(\rho) = in$ ならば $m_k = (P, Excuse(C,B))$ を選択する。
   選択できない場合は $L_{PAF_C^{d_{k+1}}}(\rho) \neq in$ であっても $m_k = (P, Excuse(C,B))$ を選択する
2. $d_k \neq d_0$ であるとき
   $\rho \in PAR_C^{d_{k+1}}$、$L_{PAF_C^{d_{k+1}}}(\rho) = in$ ならば $m_k = (P, Pass)$ を選択する。そうでなければ、$m_0 = (P, Assert(\rho,-))$ を選択する。
3. 言明
   $\rho \in PAR_C^{d_{k+1}}$、$L_{PAF_C^{d_{k+1}}}(\rho) = in$ ならば $m_k = (P, Assert(B,A))$ を選択する。
4. パス
   $m_k = (P, Pass)$ を選択する

戦略 S_F は $\mathbf{X}$ がプロトコルに従った可能手を任意に一つ選択するような戦略である。

戦略 $\mathbf{S_F}$

1. $d_k = d_0$ であるとき
   $m_0 = (X, Assert(\rho,-))$ を選択する
2. $m_{k-1} = (Y, Suspect(B,A))$ であるとき、
   $m_k = (X, Excuse(C,B))$ を選択する

3. $m_{k-1} \neq (Y, Suspect(B, A))$ であるとき、条件を満たすものが複数ある場合は任意に 1 つ選択する。
   (a) $m_k = (X, Assert(B, A))$
   (b) $m_k = (X, Suspect(B, A))$
   (c) $m_k = (X, Pass)$

戦略 S_H は **X** が議題の発言を除いて正直な発言しかしないような戦略である。それ以外は戦略 S_F と同じである。

4 評価実験

4.1 条件設定

前節で述べた対話モデルを実装することにより、以下の 3 つを調査する。

1. エージェントの戦略により対話の勝敗に違いがあるのか
2. エージェントが初期に持つ AF により対話の勝敗に違いがあるのか
3. 上記 2 つの調査から、**P** の勝率が高くなるような戦略を考案

UAF は対話モデルにおける知識のすべてを表す AF であり、実験では以下のように設定する。

設定 1 UAF は木構造であり、以下の条件を満たす。

- 根ノードが **P** の発言する議題
- ノード数は 15 個以下
- 子ノードの数は 3 個以下
- 木の幅は 4 から 6
- 木の深さは 2 から 5

設定 2 木構造をした $AF = \langle AR, AT \rangle$ において、根ノードから葉ノードまでの枝 $b_i (i = 1, \ldots, k)$ に含まれるすべてのノードの集合を $S_{AR_{b_i}}$ とする。$AR' \subseteq \{S_{AR_{b_1}}, \ldots, S_{AR_{b_k}}\}$ かつ $|AR'| < k/2$ を満たす任意の AR' に対して $S_{AR} = \{AR | AR \in AR_b, AR_b \in AR'\}$ を定義する。

S_{AR} は AF に含まれる半分未満の枝に出現するノードを集めたものである。AF の部分 AF AF_S に対して、$AF_S = \langle S_{AR}, AT_S \rangle$ とする。AF_S をその論証の数によって以下のようによぶ。$|S_{AR}| = |AR|$ のとき $all(AF)$、$|S_{AR}| = \lfloor (|AR|/2) \rfloor$ のとき $half(AF)$、$|S_{AR}| = \lfloor (|AR|/4) \rfloor$ のとき $quarter(AF)$、$|S_{AR}| = 0$ のとき $empty(AF)$ とよぶ。

実験において、最初に与える AF_P、PAF_C、AF_C、PAF_P を AF の初期状態とよぶ。$UAF1$ 個から実験の条件設定を満たす AF の初期状態をそれぞれ複数個用意し、対話をシミュレートする。

設定 3 実験で用いる AF_P、AF_C、PAF_C、PAF_P の形として All_{AF}、$Half_{AF}$、All_{PAF}、$Half_{PAF}$、$Quarter_{PAF}$、$Empty_{PAF}$ を以下のように設定する。

- $All_{AF} = all(UAF)$
- $Half_{AF} = half(UAF)$
- $All_{PAF} = all(AR_P \cup AR_C)$
- $Half_{PAF} = half(AR_P \cup AR_C)$
- $Quarter_{PAF} = quarter(AR_P \cup AR_C)$
- $Empty_{PAF} = empty(AR_P \cup AR_C)$

設定 4 実験 1、2 で用いる AF の初期状態の設定を Type1 としたときに、Type1 を表 1 のように 4 つ設定する。

表 1: 実験 1、2 で用いる AF の初期状態の設定

Type1	AF_P	PAF_C	AF_C	PAF_P
Ⅰ	All_{AF}	All_{PAF}	$Half_{AF}$	$Empty_{PAF}$
Ⅱ	All_{AF}	$Empty_{PAF}$	$Half_{AF}$	All_{PAF}
Ⅲ	$Half_{AF}$	All_{PAF}	All_{AF}	$Empty_{PAF}$
Ⅳ	$Half_{AF}$	$Empty_{PAF}$	All_{AF}	All_{PAF}

P が非常に有利な状況が Ⅰ、**P** が有利であるが **C** により不正直な発言を指摘されやすい状況が Ⅱ、**C** が有利であるが **P** により不正直な発言を指摘されやすい状況が Ⅲ、**C** が非常に有利な状況が Ⅳ となる。

実験 1 では、**P** の戦略を S_F と S_H と変えて、不正直な発言ができることで勝敗に差があるかを調査し、実験 2 では、**P** の戦略を S_F と S_Y と変えて、Yokohama の戦略の有効性を調査した。**C** の戦略はいずれも S_F に固定した。

設定 5 実験 3 で用いる AF の初期状態の設定を Type2 としたときに、Type2 を表 2 のように 4 つ設定する。

表 2: 実験 3 で用いる AF の条件設定

Type2	AF_P	PAF_C	AF_C	PAF_P
Ⅴ	All_{AF}	All_{PAF}	$Half_{AF}$	$Empty_{PAF}$
Ⅵ		$Half_{PAF}$		
Ⅶ		$Quarter_{PAF}$		
Ⅷ		$Empty_{PAF}$		

P が相手の知識を全て予測している状況が Ⅴ、半分予測している状況が Ⅵ、4 分の 1 だけ予測している状況が Ⅶ、全く予想していない状況が Ⅷ となる。

実験 3 では、PAF_C 以外を固定したときに PAF_C に含まれる論証の数を変えて、結果を比較する。**P** と **C** はいずれも戦略 S_F を用いる。

実験 1、2、3 では 50 個の UAF を用意する。UAF1 個から実験の条件設定を満たす AF の初期状態を実験 1、2 では Type1 の 4 種類、実験 3 では Type2 の 4 種類それぞれを 5 組用意する。Type1、2 ごとに設定された AF の初期状態をそれぞれ 250 パターンずつ用意する。

4.2 結果および考察

UAF と AF の初期状態を固定したものを 1 つのパターンとし、全対話をシミュレートする。各パターンにつき、全対話の中で **P** が勝利する対話の占める割合を調査する。

AF の初期状態が与えられたとき、対話をシミュレートした場合の全対話数を D_{sum} とする。全対話の内、**P** が説得に成功した対話の総数を D_{suc}、**P** が **C** の不正直な発言を見破った対話の総数を $D_{C_{lie}}$ とする。**P** の勝率を $W = (D_{suc} + D_{c_{lie}})/D_{sum} \times 100(\%)$ と定義する。**P** が戦略 S を用いたときの **P** の勝率を $W(S)$ と表す。

実験結果 1 Type1 の各 250 パターンの中で **P** の勝率を比較した結果は表 3 のようになる。

表 3: $W(S_F)$ と $W(S_H)$ の比較 (%)

TYPE1	$W(S_H) < W(S_F)$	$W(S_H) = W(S_F)$	$W(S_H) > W(S_F)$
I	35.6%	61.2%	3.2%
II	24.4%	59.6%	16%
III	0.4%	86.8%	12.8%
IV	9.6%	84%	6.4%

I のときに、**P** が S_F を用いる方が、S_H を用いるよりも勝率が高くなる。この要因は 2 つある。1 つめは **P** が不正直な発言をすることにより、反論する機会が増えて **C** の説得が成功しやすくなるからである。2 つめは **P** が不正直な発言をすることで **C** も不正直な発言が増え、その結果、**P** が **C** の不正直な発言を見破る可能性が増えるからである。

I では **C** が予測知識を持っていなかったが、II では **C** が予測知識を持っているため、**C** は **P** の不正直な発言を指摘することができる。**C** が不正直な発言を指摘することによって **P** の不正直な発言が見破れらやすくなったため、勝率が低くなる。I で述べたような勝率が高くなる要因とともに、低くなる要因があるので、結果として II では戦略の違いによる大きな差がでなかった。

実験結果 2 Type1 の各 250 パターンの中で **P** の勝率を比較した結果は表 4 のようになる。

表 4: $W(S_Y)$ と $W(S_F)$ の比較 (%)

Type1	$W(S_F) < W(S_Y)$	$W(S_F) = W(S_Y)$	$W(S_F) > W(S_Y)$
I	0%	61.2%	38.8%
II	0.4%	59.6%	40%
III	3.2%	84%	12.8%
IV	14%	84%	2%

I, II の場合ともに **P** が S_Y を用いる方が、S_F を用いるよりも勝率が低くなる。この要因は、S_Y は対話で敗北してしまうような可能手を避けることが目的で、**P** が不正直な発言を指摘できる状況であっても、不正直な発言を指摘しないからである。

実験結果 1、2 において、I、II のときは戦略によって勝率が変化する割合が約 40%あるのに対して、III、IV のときはわずか 15%しかない。また、III、IV のときでは、戦略よりも初期に与えられた AF の影響が大きい。

実験結果 3 K_1,K_2 を Type2 の V〜VIII のいずれかとする。**P** と **C** の戦略が S_F であり、K_1,K_2 であるときの **P** の勝率を $W(K_1)$,$W(K_2)$ として、$W(K_1)$ と $W(K_2)$ を比較する。**P** の勝率を比較した結果は表 5 のようになる。

表 5: $W(K_1)$ と $W(K_2)$ の比較 (%)

	$W(K_2) < W(K_1)$	$W(K_2) = W(K_1)$	$W(K_2) > W(K_1)$
	K_1 = V		
K_2 = VI	19.2%	77.2%	3.6%
K_2 = VII	16.4%	81.2%	2.4%
K_2 = VIII	11.6%	87.6%	0.8%
	K_1 = VI		
K_2 = VII	7.6%	82.4%	10%
K_2 = VIII	0.6%	82.4%	11.6%
	K_1 = VII		
K_2 = VIII	2.8%	90.8%	6.4%

対話の勝敗は AF_P と AF_C に大きく依存している。V のときと VI、VII、VIII のときを比較したときに前者の方が勝率が高くなる。また、**P** が相手の知識を全て予測している状況以外では、**P** が相手の知識を予測していたとしても勝率に与える影響はほとんどない。

V のときと VIII のときの結果を比較したときに、**C** が相手の知識を何も予測していないにもかかわらず、**P** の不正直な発言が見破られてしまう場合が見られた。この原因とこれを避けるための戦略について以下で述べる。

4.3 評価実験を基にした戦略の考案及び実験

実験結果 3 により、**C** が相手の知識を何も予測していないにもかかわらず、**P** の不正直な発言が見破られてしまう場合が見られた。その原因は **P** が *Suspect*、*Excuse* により論証を発言した後に *Assert* で同じ論証を発言するからである。これを避けるために、*Suspect*、*Excuse* により論証を発言した後に *Assert* を用いないようにする戦略 S_A を考案する。

戦略 $\mathbf{S_A}$

1. $d_k = d_0$ であるとき
   $m_0 = (X, Assert(\rho, -))$ を選択する
2. $m_{k-1} = (Y, Suspect(B, A))$ であるとき、
   $m_k = (X, Excuse(C, B))$ を選択する
3. $m_{l-1} = (Y, Suspect(B, A)/Excuse(B, A))$
   $(0 \leq l \leq k)$ であるとき
   $m_k = (X, Assert(B, A))$ を選択しない
4. $m_{k-1} \neq (Y, Suspect(B, A))$ であるとき条件を満たすものが複数ある場合は任意に一つ選択する。
   (a) $m_k = (X, Assert(B, A))$ を選択する
   (b) $m_k = (X, Suspect(B, A))$ を選択する
   (c) $m_k = (X, Pass)$ を選択する

P が新たに提案した戦略 S_A を用いるときと S_F、S_H、S_Y を用いるときの勝率を比較する。

実験 1 **P** の戦略を S_A、S_F、S_H、S_Y と変えて、結果を比較する。**C** の戦略はいずれも S_F とする。

実験 2 実験 3 と同様に、PAF_C 以外を固定したときに PAF_C に含まれる論証の数を変えて調査する。今回は **P** は戦略 S_A を用い、**C** は戦略 S_F を用いる。この結果を実験 3 のものと比較する。

実験 1,2 では 50 個の UAF を用意する。UAF1 個から実験の条件設定を満たす AF の初期状態を実験 1 では Type1 の 4 種類、実験 2 では Type2 の 4 種類それぞれを 5 組用意する。Type1,2 ごとに設定された AF の初期状態をそれぞれ 250 パターンずつ用意する。

実験結果 4 Type1 の各 250 パターンの中で **P** の勝率を比較した結果は以下のようになる。

表 6: $W(S_A)$ と $W(S_F)$ の比較 (%)

TYPE1	$W(S_F) < W(S_A)$	$W(S_F) = W(S_A)$	$W(S_F) > W(S_A)$
I	38.8%	61.2%	0%
II	28.8%	59.6%	11.6%
III	16%	84%	0%
IV	12.8%	84%	3.2%

表 7: $W(S_A)$ と $W(S_H)$ の比較 (%)

TYPE1	$W(S_H) < W(S_A)$	$W(S_H) = W(S_A)$	$W(S_H) > W(S_A)$
I	37.2%	61.2%	1.6%
II	27.6%	59.6%	12.8%
III	12%	87.2%	0.8%
IV	14%	84%	2%

表 8: $W(S_A)$ と $W(S_Y)$ の比較 (%)

TYPE1	$W(S_Y) < W(S_A)$	$W(S_Y) = W(S_A)$	$W(S_Y) > W(S_A)$
I	38.8%	61.2%	0%
II	38%	59.6%	2.4%
III	12.8%	87.2%	0%
IV	1.6%	84%	14.4%

ほぼ全ての場合で、**P** が S_A を用いる方が他の戦略を用いるよりも勝率が高くなることから、他の戦略を用いるよりも S_A を用いる方が有効である。

実験結果 5 Type2 の各 250 パターンの中で **P** の勝率を比較した結果は表 9 のようになる。

表 9: $W(S_A)$ と $W(S_F)$ の比較 (%)

Type2	$W(S_F) < W(S_A)$	$W(S_Y) = W(S_A)$	$W(S_F) > W(S_A)$
V	32.8%	67.2%	0%
VI	31.6%	68.4%	0%
VII	28.8%	71.2%	0%
VIII	26%	74%	0%

この結果から、戦略 S_A を用いると **P** の予測がどのくらいあるかにかかわらず、S_A の方が S_F よりも有効であることがわかった。

実験結果 4,5 より、**P** が S_A を用いることでほぼすべての場合において、他の戦略を用いるよりも勝率が高くなることが判明した。

5 結論

本研究では Yokohama の提案した、予測知識をもつ対話モデル上での説得対話の戦略の評価実験を行った。本研究の成果は、各エージェントに与えられた知識が相手を説得することや不正直な発言を見破ることに影響を及ぼすことを実験により明らかにしたことである。また、エージェントが不正直な発言をすることにより説得に成功する機会が増えることを明らかにした。さらに、評価実験の考察から新たな戦略を考案することにより、説得に成功する対話を増やすことができた。

今後の課題としては評価実験の結果をより詳細に調査することで、新たな性質を発見し、証明することである。また、本研究の不正直な論証の定義はラベリングに依存しているため、基礎ラベリング以外のラベリングの下で対話モデルの評価実験をする必要がある。

参考文献

[1] Amgoud, L., Maudet, N., and Parsons, S. (2000). Modeling dialogues using argumentation. In ICMAS2000, pages 31-38.

[2] Baroni, P., Caminada, M., and Giacomin, G. (2011). An introduction to argumentation semantics. The Knowledge Engineering Review, 26(4):365-410.

[3] Dung, P. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence, 77(2):321-358.

[4] Sakama, C. (2012). Dishonest arguments in debate games. In COMMA2012, pages 177-184.

[5] Takahashi, K. and Yokohama, S. (2017). On a formal treatment of deception in argumentative dialogues. In EUMAS-AT2016, Selected papers, pages 390-404.

[6] Yokohama, S. and Takahashi, K. (2016). What should an agent know not to fail in persuasion? In EUMAS-AT2015, Selected papers, pages 219-233.