論証構築のための情報探索対話戦略の最適化

勝見 久央 [1] 平岡 拓也 [2] 本浦 庄太 [2] 山本 風人 [2] 定政 邦彦 [2] 吉野 幸一郎 [1,3] 中村 哲 [1]

[1] 奈良先端科学技術大学院大学
[2] 日本電気株式会社
[3] 科学技術振興機構 さきがけ

1 はじめに

論証に基づく議論の場では，論証の根拠となる事実を収集するための情報探索対話 [12] を効率的に行うことが重要であるが，実際の議論の場で人間がこの対話戦略を最適化することは困難である．例えば，裁判では裁判官は判決の根拠となる証拠を証人等に問い合わせて収集する．この際，対話時間 (問い合わせ回数) に関する制約や，相手から期待した返答が得られない場合等を考慮しながら，主張の根拠となる適切な情報を適切な量だけ収集する必要があるが，これを可能とする対話戦略は自明でない．

自律エージェント間の論証に基づいた情報探索対話に関する研究は存在する [3, 9] が，これらの研究で提案された方策は人手で作成されたものであり，先に述べた時間的制約などを考慮した最適化は行われていない．他方，論証に基づいた議論における対話戦略の最適化に取り組んだ研究もいくつか存在する [1, 5, 10]．これらでは，あらかじめ提示可能な論証がいくつか存在する状況において，それらをどのように相手に提示していくかを決定する対話戦略の最適化に焦点が当てられている．

本研究では，論証構築のための情報探索対話をマルコフ決定過程に基づくとして定式化し，質問者の対話戦略の最適化に取り組む．本研究で扱う対話における質問者の目的は，情報探索対話においてできるだけ少ない問い合わせ回数で効率的に論証の根拠となる情報を収集することである．そこで，合理的な論証を少数回の回答者への問い合わせで構築できた場合に高い累積報酬を与える報酬関数を定義する．質問者の対話戦略は，回答者から期待通りの情報が得られないなどの可能性を考慮して，この報酬の期待値を最大化するように最適化される．

まず，2 章で論証に基づく情報探索対話の枠組みについて説明する．3 章では，2 章で説明する情報探索対話の最適化手法をマルコフ決定過程に基づき定式化する．4 章ではシミュレータを用いた最適化対話戦略の学習と，評価実験の結果について示し考察する．最後に 5 章で，本研究のまとめと，今後の課題を示す．

本研究の主たる貢献は，論証構築に基づく情報探索対話において話者の対話戦略を最適化し，話者の知識の規模がある程度大きいドメインにおいてその有効性を例証したことである．

2 論証構築のための情報探索対話

論証構築のための情報探索対話では，質問者 Q はある主張を導くための合理的な根拠の要素となり得る事実を，回答者 A から取得する (図 1)．本章では論証とその構築方法，そして論証構築を行うための情報探索対話について説明する．

論証 $\langle \Phi, \alpha \rangle$ は，ある主張 α とそれを導くための根拠 Φ の組である．根拠は推論規則，仮説から構成される．本研究では，Φ と α は一階述語論理式の集合として定義する．図 1 に,「次郎の呼気中から微量のアルコールが検出されたら，次郎は酒気帯び運転をしている．なぜなら，次郎は車を運転していたからだ.」という論証の例を示す．ここで，主張は「次郎は酒気帯び運転をしている / drunk_drivng(Jiro)」であり，根拠は「次郎は車の運転をしていた / drove_car(Jiro)」と「次郎の呼気中から微量のアルコールが検出された, / asm(detected_alcohol(Jiro))」等である．asm は仮説を表す関数である．本研究では Assumption-based argumentation[2] に基づいて仮説も根拠に含めることを許す．これは，実際の議論で頻繁にやり取りされるような，根拠に仮説を含めた論証を考慮するためである．さらに,「次郎は酒気帯び運転をしている / drunk_drivng(Jiro)」などの論理式を，取り扱う論理式の最小単位である原子論理式とする．

論証は Q の知識 K_Q と主張 α から $K_Q \cup H \models \alpha$ を満たすような仮説 H を仮説推論モデル L を用いて補完し，$\langle \Phi \equiv K_Q \cup H, \alpha \rangle$ として構築される[1]．ただし，$K_Q \models \alpha$ が成立しているときは $\langle \Phi \equiv K_Q, \alpha \rangle$ として構築される．なお，K_Q は原子論理式 $p_1 ... p_n, q$ を用いて $p_1 \wedge ... \wedge p_n \rightarrow q$ と表せる論理式の集合である．本稿では $n = 0$ のものを事実と呼び，それ以外のものを推論規則と呼ぶ．このとき，論証の合理性は仮説推論モデル L の評価関数 E を用いて評価できる．仮説推論モデル L には確率的仮説推論モデルを含む，説明の良さに対する評価関数 E を用いて仮説候補の集合 $\mathscr{H}$ から仮説 $\arg\max_{h \in \mathscr{H}} \mathrm{E}(h)$ を導くような様々なモデルが考えられる．本研究では重み付き仮

[1] ただし，α の導出に必要のない推論規則は Φ から除外される．

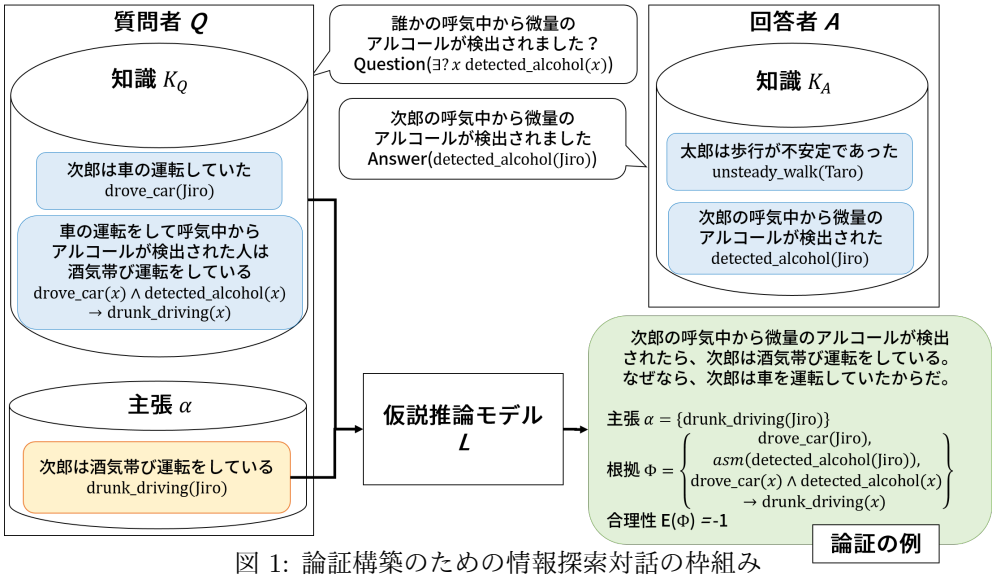

図 1: 論証構築のための情報探索対話の枠組み

説推論 [7, 8, 6] を採用し，その評価関数 E(Φ) を主張に対する根拠の合理性として用いる[2]．また，E(Φ) は重み付き仮説推論の性質から，補完された仮説数の多さや仮説の蓋然性の低さに応じて低くなる．図 1 の例では，K_Q 中の「drove_car(Jiro)」と「drove_car(x) ∧ detected_alcohol(x) → drunk_driving(x)」だけでは「drunk_driving(Jiro)」を導出できないため，「detected_alcohol(Jiro)」が仮説されて論証が構築され，合理性もその分だけ減点されている．

論証構築のための情報収集対話では，Q は E(Φ) が高い Φ を求めるために必要な事実を A から収集することを目的とする．Q は問い合わせを通じて A から事実の収集を行い，収集された事実を基に K_Q を更新する．また，A は Q から問い合わせが来た場合は，自身の知識 K_A を基に回答を行う．図 1 の例では，Q は A に「誰かの呼気中から微量のアルコールが検出されましたか？ / Question(∃?x detected_alcohol(x))」と問い合わせている．A は K_A 中に「次郎の呼気中から微量のアルコールが検出された / detected_alcohol(Jiro)」が含まれていたので,「次郎の呼気中から微量のアルコールが検出されました / Answer(detected_alcohol(Jiro)) 」と回答している．Q はこの回答に基づいて知識を更新している ($K_Q \leftarrow K_Q \cup \{\text{detected\_alcohol(Jiro)}\}$)．この知識更新により論証構築時に「detected_alcohol(Jiro) 」を仮説する必要がなくなり根拠の合理性が高くなる．このような Q の問い合わせとそれに対する A の回答は，1)E(Φ) が与えられた閾値を上回るまで，あるいは 2) このやり取りが一定回数を超えるまで，繰り返される．

Q の問い合わせ内容の候補は K_Q 中の推論規則中に含まれる原子論理式に対応する．図 1 中では，K_Q 中に「drove_car(x) ∧ detected_alcohol(x) → drunk_driving(x)」が含まれているため，Q の問い合わせ内容の候補は「∃?x drove_car(x)」,「∃?x detected_alcohol(x)」と「∃?x drunk_driving(x)」である．ただし，本研究では，推論規則については最初から Q は全て知っているものとして K_Q に含まれており，問い合わせの候補としない．他方，A は K_A に含まれる原子論理式のみを回答内容に含めることとする．すなわち，A は虚偽の内容を回答する事は出来ないことを仮定している．

3 論証構築のための情報探索対話戦略の最適化

本章では，マルコフ決定過程に基づいた論証構築のための情報収集対話のモデル化について説明する．なお，時刻 t におけるエージェントのアクション，状態，報酬をそれぞれ a_t, s_t, r_t とする．また，t における Q の知識を $K_{Q,t}$ とする．

アクション a_t は Q が A に問い合わせる内容を表す．Q はあらかじめ問い合わせ内容の候補の辞書を保持しており，a_t は問い合わせる内容のインデックスとなる．例えば，問い合わせ内容の候補の辞書として [∃?x drove_car(x) : 0, ∃?x detected_alcohol(x) : 1] を保持している場合は，$a_t \in \{0, 1\}$ である．

状態 s_t は，Q が過去に A に問い合わせた内容の履歴と，E(Φ_t) を表す．Φ_t は $K_{Q,t}$ を用いて L によって求められた根拠であり，Q が過去に問い合わせた内容に対応した要素が 1 となる 2 値ベクトルを用いて表現される．このベクトルと E(Φ) の値の結合ベクトルによって状態が表現される．

報酬 r_t は回答者からの情報獲得による根拠の合理性の改善度合いに応じて与えられる．論証構築のための

[2] より具体的な合理性の算出方法については [7] の Cost を参照されたい．本研究では E(Φ) は Cost に-1 を乗じて算出される．

情報探索対話における最適化対話戦略は，最小のアクション選択回数で，主張のより合理的な根拠を求めることが出来る情報を収集することである．本報酬はそのような最適化対話戦略を学習するように定義した．具体的には，r_t は $\mathrm{E}(\Phi_{t+1}) - \mathrm{E}(\Phi_t) - c$ として定義される．ここで c は時間圧力となるアクションごとのペナルティをあらわす定数である．$\mathrm{E}(\Phi)$ は前章で述べた Φ の合理性の評価値である．Φ_t は $K_{Q,t}$ から，また Φ_{t+1} は $K_{Q,t+1}$ から，それぞれ L によって求められた根拠である．

本研究では，Q の対話戦略を最適化する際に A のシミュレータを用いる．シミュレータは Q からの問い合わせに対して，問い合わせ内容を知っていればそれに対応する事実を返し，知らなければ何も返さない．従ってシミュレータが知っている事実は，シミュレーション開始時に与えられる K_A に応じて決定される．

4 評価実験

3 章で述べたモデルに対して，強化学習とモンテカルロ木探索を用いて Q の対話戦略の最適化とその評価を行った．評価では最適化時と異なる K_A に従う A と対話を行い，平均累積報酬に基づいて対話戦略の優劣を評価した．また比較手法として，人手で作成した対話戦略とランダムにアクション選択する対話戦略を用意した．

4.1 話者の知識構築とデータ

実験において A のシミュレータに K_A を，Q のシミュレータに K_Q をそれぞれ与える．K_A にはロールプレイングに基づいて人手で作成された独占禁止法に違反する電子メールのやり取りに対して，Chapas [3] を用いた述語項構造解析，Zunda [4] を用いたモダリティ解析を行って得られた原子論理式の集合を使用した．例えば，ロールプレイングにおいて，A 社の太郎がその競合会社の B 社の次郎に送信したメールに「価格についてお知らせいたします.」という一文が含まれているとき，これは「provide_price_infomation(太郎, 次郎)」という原子論理式に変換される．これらによって 5〜10 通の電子メールからなる 1 回のやり取りが 100〜600 個の原子論理式の集合に変換され，K_A として与えられた．さらに，同様の処理を 100 回の異なるやり取りに対して適用し対応する 100 種類の K_A が用意された．そのうち 50 種類は対話戦略の学習に，残りの 50 種類は評価に利用した．

また，K_Q には電子メール上での犯行を審査するための推論規則 (例えば，「競合会社同士の人間が入札価格情報のやり取りをするとカルテルである．/ competitive(x, y) ∧ provide_price_infomation(x, y) → cartel(x, y)」) を与えた．推論規則は約 200 種類あり，それらの規則に含まれる原子論理式 (すなわち Q の問い合わせ内容の候補) は約 160 種類ある．

本実験設定では，Q は「電子メール上の容疑者 S1 と容疑者 S2 のやりとりはカルテルに該当する」を主張する論証 $\langle \Phi, \mathrm{cartel}(\text{容疑者 S1，容疑者 S2}) \rangle$ の合理性 $\mathrm{E}(\Phi)$ が出来るだけ高くなるよう K_A に含まれる原子論理式を A から収集する．また，Q は事実を表す 100〜600 個の論理式の集合 K_A から，合理的な根拠を求めるために必要な平均 10 個程度の論理式を出来る限り無駄なく収集する必要がある．この最適化対話戦略は，各対話開始時にはアクションの選択候補が約 160 種類存在し，その選択を最長で 50 回まで行うと約 $160 \times 159 \times 158 \times ... \times 111 \approx 3 \times 10^{106}$ 通りの膨大な探索空間から学習を行う必要がある．

4.2 比較手法

対話戦略評価のために，次の 4 種類の対話戦略を用意した：

RL: モデルフリー強化学習の一種である Double DQN[11] を用いて最適化された対話戦略．Double DQN では，状態 s_t におけるアクション a_t の評価関数である行動価値観数 (Q 関数; $Q(s_t, a_t)$) について，ニューラルネットワークを用いた関数近似によって最適 Q 関数 Q^* を学習する．このとき，ネットワークの教師信号は，

$$r_{t+1} + \gamma Q(s_{t+1}, \arg\max_a Q(s_{t+1}, a; \theta_t); \theta_t^-)$$

として計算される．ただし，θ_t，θ_t^- は，2 種類のネットワークのパラメータ，γ は報酬の割引率である．割引率が小さいほど，より即時的な報酬を優先して最適 Q 関数を学習する．状態 s_t における最適なアクションは，$\arg\max_a Q^*(s_t, a)$ として得られる．前述の A のシミュレータをエピソード毎にランダムに切り替え，合計 1000 エピソード対話を通じて対話戦略を最適化した．実験では割引率 γ を 0.95，0.5 とした $\mathrm{RL}(\gamma = 0.95)$，$\mathrm{RL}(\gamma = 0.5)$ の 2 種類を用意した．

MCTS: モンテカルロ木探索の一種である Ensemble-UCT[4] の変種を用いて最適化された対話戦略．50 種類の異なる K_A に従った最適化用の A のシミュレータそれぞれに対して最適なアクション系列を求め，それらの多数決を取って最適化対話戦略とした．各 UCT で 1 回のアクション選択に対し 80 回のロールアウトによってアクションを選択して得られた MCTS (80) と，100 回のロールアウトによってアクションを選択して得られた MCTS (100) を用意した．

Handcrafted: 人手で作成した対話戦略．K_Q 中の推論規則を用いて，α に対応する論理式に対して後ろ向き連鎖を行い，少数回の推論規則の適用によって到達する論理式ほど優先して A に問い合わせる．すなわち，推論規則の適用回数を

[3] https://sites.google.com/site/yotarow/chapas
[4] https://jmizuno.github.io/zunda/

表 1: 強化学習と比較手法のテスト結果

手法	RL (γ=0.95)	RL (γ=0.5)	MCTS (80)	MCTS (100)	Handcrafted	Random
平均累積報酬	-0.0831	0.1765	0.3180	0.3199	-0.2180	-0.3056

深さとするような幅優先探索を行う．例えば，α が C(X) で，K_Q が「A(x) $\wedge$ B(x) $\rightarrow$ C(x)」，「D(x) $\wedge$ E(x) $\rightarrow$ A(x)」の 2 つで構成されているとき，C(x)，A(x)，B(x)，D(x)，E(x) のような順で問い合わせる．ただし，A(x) と B(x)，D(x) と E(x) のように推論規則の適用回数が同数の論理式（すなわち，同じ深さの論理式）は，ランダムに問い合わせ順序を決定した．

Random: ランダムにアクション選択する対話戦略．

4.3 評価

各対話戦略の評価結果を表 1 に示す．この結果から，強化学習とモンテカルロ木探索による最適化対話戦略が，人手で作成した対話戦略とランダムな対話戦略に比べて高い性能 (平均累積報酬) を達成していることがわかる．Handcrafted の対話戦略が強化学習やモンテカルロ木探索の手法に劣っているのは，各対話ごとに変化する回答者の知識 K_A を考慮せず，推論規則の情報のみを用いているからと考えられる．また，報酬として根拠の評価関数である E(Φ) を使用しているが，仮説推論の性質上，根拠となるリテラルが一定数揃わないと E(Φ) が増加しない．こうしたケースにおいては，強化学習よりもモンテカルロ木探索の方が，ロールアウトによって正確に報酬を見積もることができるため，より最適な対話戦略を学習できる．このため，モンテカルロ木探索の手法が強化学習に比べてテスト結果において勝っていると考えられる．すなわち，MCTS に従った質問者が，論証構築に十分な数と質の情報を収集できており，問い合わせ回数も遥かに少なく済んでいる．

図 2 は RL の 1000 エピソードまでの学習曲線である．この図から RL($\gamma = 0.5$)，RL($\gamma = 0.95$) ともに一定の平均累積報酬で学習が収束していることがわかる．また，割引率 γ を大きくして学習された方策 (RL($\gamma = 0.5$)) の方が，各エピソードにおける累積報酬が高いことがわかる．これは，例えば，主張である drunk_driving(Jiro) の根拠になりうる detected_alcohol(Jiro) と drove_car(Jiro) に対して，問い合わせ順序が合理性に影響しないため，即時報酬を優先して重要そうな事実を次々に問い合わせていく貪欲な対話戦略のほうが有利となったためと考えられる．

5 おわりに

本研究では，合理的な論証を構築するための事実を収集する情報探索対話をマルコフ決定過程で定式化し，強化学習やモンテカルロ木探索による最適化対話戦略が有効であることを示した．しかし，これらは実際の

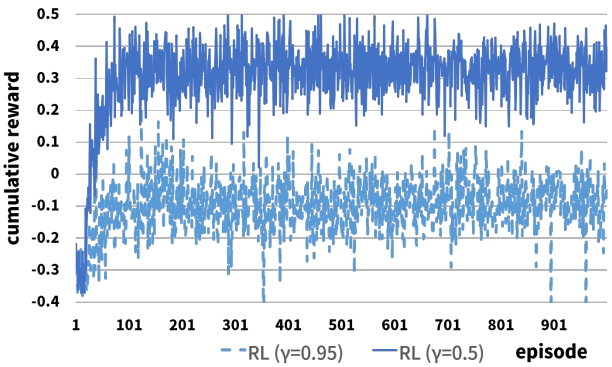

図 2: 強化学習での学習曲線

人間との対話において頻繁に発生する言語理解誤りや，相手が質問者の問い合わせに対して嘘をつく場合を考慮できていない．また，本研究では回答者は事実に相当する原子論理式のみを質問者に提示したが，それに加えて推論規則等の，より多様な情報を提示するような設定にも拡張出来る．さらに，4.3 節で述べたとおり，本研究で使用した報酬関数は疎かつ逐次的であり，改善の余地がある．今後はこれらも考慮した最適化対話戦略の作成，評価実験を行う．

参考文献

[1] Sultan Alahmari, Tommy Yuan, and Daniel Kudenko. Reinforcement learning for abstract argumentation: A q-learning approach. In *Adaptive and Learning Agents workshop (at AAMAS 2017)*, 2017.

[2] Phan Minh Dung, Robert A Kowalski, and Francesca Toni. Assumption-based argumentation., 2009.

[3] Xiuyi Fan and Francesca Toni. Mechanism design for argumentation-based information-seeking and inquiry. In *International Conference on Principles and Practice of Multi-Agent Systems*, pp. 519–527. Springer, 2015.

[4] Alan Fern and Paul Lewis. Ensemble monte-carlo planning: An empirical study. In *ICAPS*, 2011.

[5] Emmanuel Hadoux, Aurélie Beynier, Nicolas Maudet, Paul Weng, and Anthony Hunter. Optimization of probabilistic argumentation with markov decision models. In *IJCAI*, pp. 2004–2010, 2015.

[6] Jerry R Hobbs, Mark E Stickel, Douglas E Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, Vol. 63, No. 1-2, pp. 69–142, 1993.

[7] Naoya Inoue and Kentaro Inui. Ilp-based reasoning for weighted abduction. In *Plan, Activity, and Intent Recognition*, pp. 25–32, 2011.

[8] Ekaterina Ovchinnikova, Niloofar Montazeri, Theodore Alexandrov, Jerry R Hobbs, Michael C McCord, and Rutu Mulkar-Mehta. Abductive reasoning with a large knowledge base for discourse processing. In *Computing Meaning*, pp. 107–127. Springer, 2014.

[9] Simon Parsons, Michael Wooldridge, and Leila Amgoud. An analysis of formal inter-agent dialogues. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pp. 394–401. ACM, 2002.

[10] Ariel Rosenfeld and Sarit Kraus. Strategical argumentative agent for human persuasion. In *ECAI*, pp. 320–328, 2016.

[11] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, pp. 2094–2100, 2016.

[12] Douglas Walton and Erik CW Krabbe. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press, 1995.