マルチエージェント強化学習に基づく 共同作業を自律的に行う対話システムの最適化

市川拓茉　東中竜一郎
名古屋大学大学院情報学研究科
{ichikawa.takuma.w0@s.mail,higashinaka@i}.nagoya-u.ac.jp

概要

対話で行われる複雑な共同作業では，各エージェントが自律的に行動することが望まれる．しかし，従来手法では，エージェントは相手の発話に応答するようにモデル化されており，自律性に限界がある．本研究では，複雑な共同作業を自律的に行う対話エージェントの実現を目指し，マルチエージェント強化学習を用いた対話システムの学習手法を提案する．具体的には，相手の反応が得られなくても自律的に発話するための工夫として skip トークンを導入し，2 体のエージェントに説得対話を実施させる．そして，相手が発話しない状況であっても，説得に最適な発話を行うように各エージェントを強化学習によって更新する．実験の結果，自然性や自律性に課題は残るものの，マルチエージェント強化学習によって，エージェントが自律的に説得的な発話が可能なことを確認した．

1　はじめに

近年，対話システムの普及に伴い [1, 2]，より高度な対話システムの実現を目指して，人間と協調的にタスクを遂行する対話システムの研究が盛んである [3, 4, 5]．対話で行われる複雑な共同作業では，各エージェントが自律的に行動することが望まれる．しかし，従来手法では，エージェントは相手の発話に応答するようにモデル化されており，自律性に限界がある [6, 7, 8]．

本研究では，複雑な共同作業を自律的に行う対話エージェントの実現を目指し，マルチエージェント強化学習を用いた対話システムの学習手法を提案する．具体的には，相手の反応が得られなくても自律的に発話するための工夫として skip トークンを導入し，2 体のエージェントに共同作業対話を実施させる．ここで，共同作業として，協調的に対話を行うように再設計された説得対話を扱う．対話を実施させたのち，相手が発話しない状況であっても説得に最適な発話を行うように，各エージェントを強化学習によって更新する．図 1 に提案手法により得られた対話例を示す．実験の結果，自然性や自律性に課題は残るものの，マルチエージェント強化学習によって，エージェントが自律的に説得的な発話が可能なことを確認した．

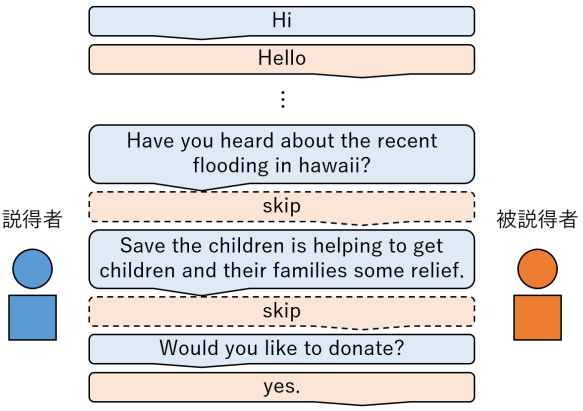

図 1: 提案手法によって得られる対話例．相手が発話しなくても（skip をしても）説得的な発話が可能．

2　関連研究

本研究は，共同作業における対話システムの研究と関連している．オンラインゲームである Minecraft において共同して建物を作成するシステム [5, 7] やユーザと一緒に物語・詩等を作成するシステム [9, 8] 等，数多くの研究が存在する．しかし，これらの研究では，エージェントは相手の発話に応答するようにモデル化されており [10, 11]，自律的に発話することができない．本研究では，行動しないことを示す skip トークンを導入し，自律的に発話を行うように各エージェントをマルチエージェント強化学習によって最適化する．

本研究は，説得対話における強化学習の研究と関

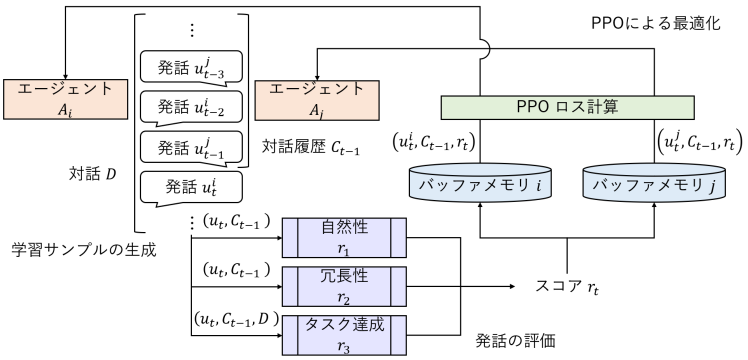

図 2: マルチエージェント強化学習の流れ．事前に定められた数の学習サンプルを生成したのち，対話に対する評価をもとに報酬を計算し，PPO アルゴリズムに基づいてエージェントを更新する．

連している．Shi ら [12] は，生成された発話について，予測された対話戦略をもとに報酬を与えて強化学習を行う手法を提案した．Samad ら [13] は，発話に対する対話戦略および感情ラベルをもとに報酬を計算し，強化学習を行う手法を提案した．これらの研究は，タスク達成（説得の結果，被説得者が購入・寄付等を行うこと）の向上のために強化学習を用いているが，いずれも発話単位でのアノテーションをもとに報酬を計算している．本研究は，対話全体に対する評価をもとに各発話の報酬を計算しており，アノテーションが不要という特徴がある．

3 提案手法

本研究では，複雑な共同作業を自律的に行う対話エージェントの実現を目指し，skip トークンの導入およびマルチエージェント強化学習を用いた学習手法を提案する．ここでは，skip トークンおよびマルチエージェント強化学習について述べる．

3.1 skip トークンの導入

従来手法では，エージェントは相手の発話に応答するようにモデル化されているため，相手の発話に依存した行動をとる必要があり，自律性に限界がある．そこで，相手の反応が得られなくても自律的に発話するための工夫として，行動しないことを示す skip トークンを導入する．具体的には，ターン t におけるエージェント A_i の発話 u^i_t を一文からなる発話文もしくは skip トークンと定義する．これにより，ターン制対話においても相手の発話が得られない状況を表現できる．本手法では，模倣学習および強化学習によって，各エージェントが相手の反応に依らずに行動できるよう最適化する．

3.2 マルチエージェント強化学習

skip トークンの導入により，エージェントは相手の反応が得られなくても発話できる．しかし，模倣学習のみでは，エージェントは対話履歴の次に続く可能性が高い発話を行うだけであり，説得に最適な発話を行うためには必ずしも十分でない．そこで，本研究では，相手が発話しないような状況であっても，エージェントが説得に最適な発話を行うようにマルチエージェント強化学習による最適化を行う．図 2 に本研究で提案するマルチエージェント強化学習の流れを示す．以降，具体的な流れを説明する．

3.2.1 学習サンプルの生成

2 体のエージェント A_i および A_j が交互に発話を行い，対話 $D = \{u^i_1, u^j_2, u^i_3, u^j_4, ..., u^i_{T-1}, u^j_T\}$ を得る．ここで，u^i_t はターン t における A_i の発話であり，T は対話の最大ターン数である．各エージェントは言語モデル $p^i_{\theta_i}$ および $p^j_{\theta_j}$ で構成される．p_θ は対話履歴 $C_{t-1} = \{u^i_1, u^j_2, u^i_3, u^j_4, ..., u^i_{t-2}, u^j_{t-1}\}$ を入力として次発話 u_t を生成する．ここで，u_t は一文からなる発話文もしくは skip トークンである．なお，θ は最尤推定によって模倣学習したものを使用する．

3.2.2 発話の評価

生成されたサンプル (C_{t-1}, u_t) に対するスコア r_t は自然性 r_1，冗長性 r_2，タスク達成 r_3 の線形和として計算される．

エージェントが相手の発話に対して自然に応答できるように自然性を報酬に加える．自然性は次発話予測モデルによって判断され，与えられた対話履歴 C_{t-1} に対して次発話 u_t が自然である場合は $r_1 = 1$，自然でない場合は $r_1 = -1$ のスコアが与えられる．

相手が発話しない状況でエージェントが同じ発話を繰り返さないように，冗長性のペナルティを報酬に加える．冗長性は C_{t-1} および u_t をもとに，単語の n-gram の重複に基づいて以下の式で定義される．

$$\mathrm{Rep}(C_{t-1}, u_t^i) = \max_{u_\tau^i \in C_{t-1}} \left(\frac{\mathrm{Ngram}(u_\tau^i) \cap \mathrm{Ngram}(u_t^i)}{\mathrm{Ngram}(u_t^i)} \right) \quad (1)$$

ここで，$\mathrm{Ngram}(u)$ は u の n-gram の集合である．上記をもとに，r_2 は以下の式で計算される．

$$r_2 = 1 - 2\mathrm{Rep}(C_{t-1}, u_t) \qquad (-1 \le r_2 \le 1) \quad (2)$$

複雑な共同作業を対象にした強化学習では，重要な行動（説得対話では，他のユーザが寄付したかどうかの情報提供等）を知るために対話行為を定義しアノテーションを行うことがあるが，このようなアノテーションには専門性が必要でありコストが大きい．そのため，共同作業全体の評価は可能であっても，個々の発話に対する評価が困難であることが多い．そこで，本研究では，対話に対する評価をもとに発話単位のスコアを計算する．タスク達成は対話分類モデルにより判断され，タスク達成の場合は $\hat{r_3} = 1$，そうでない場合は $\hat{r_3} = -1$ のスコアが基準として与えられる．発話単位のスコアは以下の式で定義される寄与率に基づいて計算される．

$$w_t^i = \mathrm{Prob}(D) - \mathrm{Prob}(D \setminus \{u_t^i, u_{t+1}^j\}) \quad (3)$$

ここで，$\mathrm{Prob}(D)$ は発話の集合である対話 D を対話分類モデルによって分類した際の予測確率である．また，r_3 は以下の式で計算される．

$$r_3 = \frac{w_t}{\max_{\tau \le T}(|w_\tau|)} \hat{r_3} \qquad (-1 \le r_3 \le 1) \quad (4)$$

なお，スコアの安定化のため，各対話における寄与率の絶対値の最大値で割ることで正規化する．

評価終了後，学習サンプル (C_{t-1}, u_t, r_t) を得る．ここで，r_t は以下の式で定義される．

$$r_t = \alpha_1 r_1 + \alpha_2 r_2 + \alpha_3 r_3 \quad (5)$$

3.2.3 PPO による最適化

事前に定められた数（ホライゾンと呼ぶ）の学習サンプルを得たのち，各エージェントを強化学習によって更新する．強化学習アルゴリズムとしては，方策ベースの手法であり学習の安定性も高いという理由から，Proximal Policy Optimization（PPO）[14] を用いた．エージェントごとに得られた学習サンプルを使用し，事前に決められたエポック数だけ PPO アルゴリズムに基づいて更新を行う．以上を 1 回の学習ループとし，定められた回数だけ繰り返す．

4 実験

4.1 データセット

本研究では，説得対話のデータセットである PersuasionForGood [15] を用いて提案手法の有効性を検証した．PersuasionForGood は慈善団体への寄付を説得する説得者と被説得者との対話コーパスであり，1,017 対話，20,932 発話が含まれる．

PersuasionForGood で扱われている説得対話では，説得者が被説得者に寄付をするよう説得を行う．本研究では，PersuasionForGood を用いた一般的な研究と異なり [12, 13]，説得者と被説得者が互いの意図を持ちつつも協調的に対話を行うタスクとして再設計した．具体的には，被説得者が寄付をすることを説得者，被説得者双方のタスク達成条件とした．

本研究では，相手の反応が得られなくても自律的に発話するための工夫として skip トークンを導入し，データセットを拡張している．そのため，複数文からなる発話は一文ごとに分解し，話者が連続する箇所に skip トークンを挿入した．

4.2 実験設定

マルチエージェント強化学習において，発話の評価に用いる次発話予測モデルおよび対話分類モデルには RoBERTa [16] を使用した．RoBERTa はテキストの分類タスク等に使用されるエンコーダモデルである．事前学習済みモデル[1)]を PersuasionForGood でそれぞれファインチューニングした．学習時の各種ハイパーパラメータとしては，バッチサイズは 64，Optimizer は AdamW，学習率は 2e–05 に設定した．損失関数には Cross Entropy Loss を使用した．

エージェントには，言語モデルとして GPT-J [17] を使用した．GPT-J は大規模言語モデルである GPT-3 [18] のクローンとして開発・公開されているテキスト生成モデルである．事前学習済みモデル[2)]を skip トークンにより拡張された PersuasionForGood で模倣学習および強化学習した．模倣学習時の各種ハイパーパラメータとしては，バッチサイズは 64，Optimizer は AdamW，学習率は 2e–05 に設定した．損失関数には Cross Entropy Loss を使用した．

強化学習時の発話生成には Nucleus Sampling を使用し，最大発話長は 36，temperature は 1.0，top-p は

1） https://huggingface.co/roberta-large
2） https://huggingface.co/hivemind/gpt-j-6B-8bit

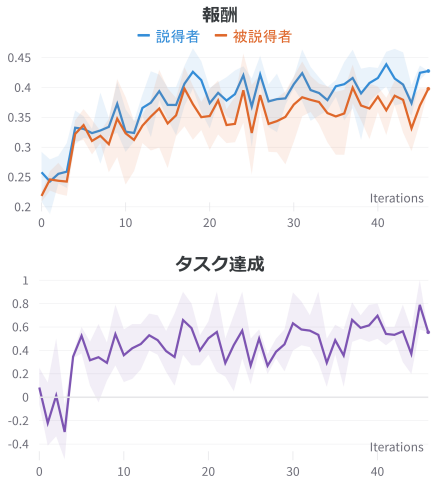

図 3: 強化学習における報酬とタスク達成のスコアの推移．実線は 3 回の試行の平均を表す．

1.0 とした．また，対話のターン数は 54 とした．冗長性の算出には 1-gram を使用し，報酬計算時の重みはタスク達成（被説得者が寄付をすること）の向上を重視して $(\alpha_1, \alpha_2, \alpha_3) = (0.25, 0.25, 0.50)$ とした．PPO の各種ハイパーパラメータとしては，ホライゾンは 256，エポック数は 4，学習ループの回数は 48，学習率は 1e–06 に設定した．公正な評価を行うため，異なる 3 種類のランダムシードで実験を行い，3 回の試行で報酬が最も高いモデルを評価に用いた．

4.3 学習結果

図 3 に強化学習における報酬とタスク達成のスコアの推移を示す．説得者，被説得者ともに，学習が進むにつれて報酬が増加し，最終的に高い値で収束した．このことから，本手法によって個々のエージェントの報酬を同時に最大化できることが確認できた．タスク達成では，学習の初めは値が 0 付近に集まっており，模倣学習のみのモデルではタスク達成が不安定であることがわかる．一方，学習が進むにつれて値が大きくなり，最終的に高い値で収束した．このことから，マルチエージェント強化学習によって，エージェントが自律的に説得的な発話が可能なことを確認した．

4.4 人手評価

人手による評価を行うために，クラウドソーシングを用いた主観評価を実施した．具体的には，以下の 2 つのモデル，1）MLE：skip トークンにより拡張された PersuasionForGood で模倣学習したモデル，2）MARL：提案手法を用いて学習したモデル，が生成した対話について人手による主観評価を実施した．

表 1: 生成された対話に対する主観評価結果（選ばれた割合）．太字は各項目における最高スコアを示す．

項目	MLE	MARL
対話の自然性	**55.56**	44.44
話者の自律性	**52.99**	47.01
対話の説得性	45.30	**54.70**

2 つのパターンについてそれぞれ 20 種類の対話を生成し，MLE と MARL の対話のペアを計 120 ペア作成した．クラウドソーシング[3)]を用いて，計 40 人の作業者が評価を行った．作業者は MLE と MARL の 2 つの対話を比較し，対話の自然性，話者の自律性，対話の説得性に関する質問それぞれについて，どちらの対話がより当てはまるか評価した．

表 1 に評価の結果（それぞれの対話が選ばれた割合）を示す．説得性については MARL の方が値が大きいことから，マルチエージェント強化学習によって，エージェントが説得に最適な発話を行うことができることが確認できた．一方，自然性と自律性については MLE の方が値が大きいことから，本手法によって自然性や自律性を完全に担保できるわけではないことも分かった．特に自然性については，強化学習時に報酬が増加したものの，報酬計算時に直前の発話のみを使用したことで長期的な自然性を考慮できなかったと考えられる．より適切な報酬設定は今後の課題である．

5 おわりに

本研究では，複雑な共同作業を自律的に行う対話エージェントの実現を目指し，skip トークンの導入およびマルチエージェント強化学習を用いた対話システムの学習手法を提案した．PersuasionForGood で扱われる説得対話を対象にマルチエージェント強化学習を行った結果，本手法によって個々のエージェントの報酬を同時に最大化できることを確認した．また，実験の結果，自然性や自律性に課題は残るものの，マルチエージェント強化学習によってエージェントが自律的に説得に最適な発話が可能なことを確認した．

本研究では，話者の双方に役割がある共同作業に着目したが，今後は双方が等しい立場で行うような，より創造的な共同作業（キャッチコピーの作成 [19]，Minecraft での建物の作成 [20] 等）に本手法を拡張したい．

3） https://www.mturk.com/

謝辞

本研究は科研費「モジュール連動に基づく対話システム基盤技術の構築」(課題番号 19H05692) の支援を受けた．また，本研究は名古屋大学のスーパーコンピュータ「不老」の一般利用を利用して実施した．

参考文献

[1] 東中竜一郎, 稲葉通将, 水上雅博. Python でつくる対話システム. オーム社, 2020.

[2] 中野幹生, 駒谷和範, 船越孝太郎, 中野有紀子, 奥村学（監修）. 対話システム（自然言語処理シリーズ）. コロナ社, 2015.

[3] Charles Rich, Candace L. Sidner, and Neal Lesh. Collagen: Applying Collaborative Discourse Theory to Human-Computer Interaction. **AI Magazine**, Vol. 22, No. 4, p. 15, 2001.

[4] Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. Data-driven models for timing feedback responses in a Map Task dialogue system. **Computer Speech & Language**, Vol. 28, No. 4, pp. 903–922, 2014.

[5] Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative Dialogue in Minecraft. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5405–5415, 2019.

[6] Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 6495–6513, 2019.

[7] Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. Learning to execute instructions in a Minecraft dialogue. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2589–2602, 2020.

[8] Tuhin Chakrabarty, Vishakh Padmakumar, and He He. Help me write a poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing. **arXiv preprint arXiv:2210.13669**, 2022.

[9] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. Wordcraft: a Human-AI Collaborative Editor for Story Writing. **arXiv preprint arXiv:2107.07430**, 2021.

[10] He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1766–1776, 2017.

[11] Daniel Fried, Justin Chiu, and Dan Klein. Reference-Centric Models for Grounded Collaborative Dialogue. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 2130–2147, 2021.

[12] Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. Refine and Imitate: Reducing Repetition and Inconsistency in Persuasion Dialogues via Reinforcement Learning and Human Demonstration. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 3478–3492, 2021.

[13] Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. Empathetic Persuasion: Reinforcing Empathy and Persuasiveness in Dialogue Systems. In **Findings of the Association for Computational Linguistics: NAACL 2022**, pp. 844–856, 2022.

[14] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. **arXiv preprint arXiv:1707.06347**, 2017.

[15] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5635–5649, 2019.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv preprint arXiv:1907.11692**, 2019.

[17] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, 2021.

[18] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In **Proceedings of the 34th Conference on Neural Information Processing Systems**, pp. 1877–1901, 2020.

[19] 周旭琳, 市川拓茉, 東中竜一郎. キャッチコピー共同作成タスクにおける対話の収集と分析. 人工知能学会全国大会論文集 第 36 回全国大会, pp. 2A6GS603–2A6GS603, 2022.

[20] 市川拓茉, 東中竜一郎. Minecraft での人間同士の共同作業における対話の分析. 言語処理学会 第 28 回年次大会 発表論文集, pp. 568–572, 2022.