URL　https://journal.ntt.co.jp/article/36100
DOI　https://doi.org/10.60249/25095002

AIコンステレーション® 実現に向けた研究開発

「AIコンステレーション®」をはじめとするマルチエージェントシステム議論では，複数の大規模言語モデル（LLM：Large Language Models）を活用することで，多様な視点に基づく課題解決が期待されています．一方，単純に言語モデルを組み合わせるだけでは，発言が類似・抽象的となり，議論が深まらないことが課題となっています．本稿では，こうした課題に対して，議論の流れを維持しつつ新たな観点を自動選択する「LLMエージェント発言多様化技術」について紹介します．

キーワード：#AIコンステレーション，#大規模言語モデル，#次世代AI

孫　晶鈺（そん しょうぎょく）／井手　綾乃（いで あやの）
吉田　司（よしだ つかさ）／渡邊　千紘（わたなべ ちひろ）
豊田　真智子（とよだ まちこ）／竹内　亨（たけうち すすむ）
NTTコンピュータ&データサイエンス研究所

複雑な社会課題とAI議論技術への期待

近年，LLM（Large Language Model：大規模言語モデル）*1の性能向上に伴い，複数のAI（人工知能）がそれぞれ異なる視点から議論に参加する，MAS（Multi-Agent System：マルチエージェントシステム）*2を用いた議論支援に大きな注目が集まっています[(1),(2)]．これらの技術は，従来であれば専門家どうしの熟議が求められる，明確な解が存在しない問題に対し，人間だけでは見落としがちな課題や解決策を提示できる可能性があるだけでなく，気付きにくい前提や潜在ニーズを見つける可能性も秘めています．さらに，時間や場所に制約されない議論環境を整えられることから，少人数では出にくい視点や，利害関係者が一堂に会することが難しい状況においても活用が期待されます．例えば，介護や防災といった地域課題を扱う複数人が参加するワークショップでは，立場によって解が異なる複雑な問題を扱っているため，従来だと現場に依存してきたノウハウに新たな切り口を与え，議論を活性化させる手段として大きな可能性が期待されます．

今後，このような人参加型ワークショップの場において，人間どうしの議論のような自然で多様な意見が交わされることに加え，明確な解が存在しない問題に対し，参加者に新たな気付きをもたらすようなAI議論技術が一層求められていきます．本稿では，その実現に向けた「LLMエージェント発言多様化技術」のアプローチを紹介し，課題整理，詳細な構成と評価，さらに今後の展開について述べます．

技術的課題：MAS議論の自然性・多様性の限界

MASにおいては，LLMの各エージェントに異なる役割や専門性を与えることで，多様な視点による議論を提供します．しかし，LLMエージェントどうしを単純に組み合わせただけの議論では，**図1**に示すようないくつかの技術的な課題が存在します．

1番目の課題は，各エージェントが議論を通じて自らの意見を修正したり発展させたりすることが難しく，結果として発言内容が類似しやすいことです．LLMエージェントは与えられた議題や役割，専門性といった指示を忠実に守る傾向が強く，一度意見を出すと，その後の議論の流れや他者の発言に応じて意見を変化させることはあまりありません．そのため，議論を繰り返しても各エージェントが似たような意見を述べてしまい，MASに期待される視点や発想の広がりが十分に生まれません．とりわけ議論に新たな価値を生み出すには，単に繰り返しではなく，類似しない多様な発言が生成されることが重要となっています．

2番目の課題は，議論を進めても論点が交差しない場合があることです．LLMは，他のエージェントの発言を参照することは可能ですが，各エージェントが与えられた指示に従う傾向があることから，自ら他者の発言に踏み込む応答はあまり行いません．そのため，発言内容が独立しがちで議論が噛み合わず，意見の対立や補完といった相互作用が生じにくくなります．特に，人間による議論で自然に見られる「他者への反論」や「共感を契機とした議論内容の深掘り」などの要素が欠け，発言が平行線のままで終わってしまう傾向があります．

＊1　LLM：大量のテキストを学習した自然言語処理モデルで，文章生成や要約が可能です．
＊2　MAS：複数のエージェントが相互にやり取りしながら課題解決に取り組む枠組み．

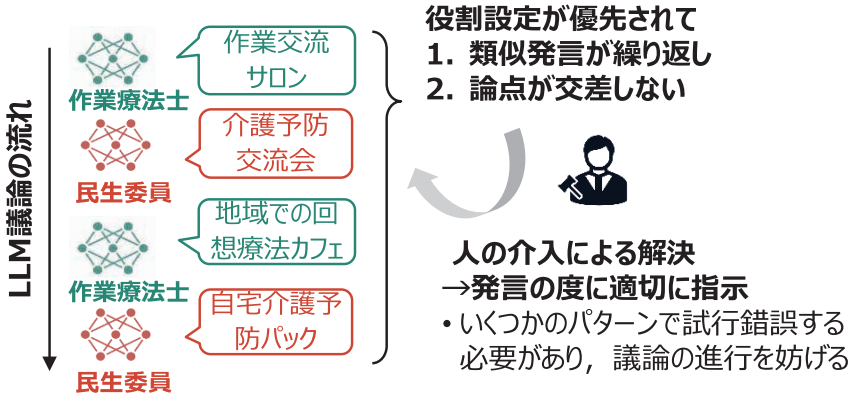

図1　MASによるLLM議論の課題

これら2つの課題はLLMの特性から自然に生じるものであり，そのままでは人参加型ワークショップにおける議論として体験価値を著しく低下させてしまいます．そのため，従来は議論の状況に合わせて人手でプロンプト*3を調整し，適切な切り口や視点を都度指示することで，類似した発言を抑止し多様な発言を保ちつつ，具体性のある議論を提供する必要がありました．しかし，このときいくつかのパターンで試行錯誤する必要があることから，運用上の負担が大きいだけでなく，ワークショップ自体の進行を妨げるため現実的な対応とはいえませんでした．

このような背景から，発言の多様性と議論の流れを自然にすることを考慮した自動制御技術が求められていました．

LLMエージェント発言多様化技術

■本技術のねらいと基本方針

前述のとおり，従来のMASにおけるLLMどうしの議論では，あらかじめ設定された議題や役割といった指示に従う性質が強く働き，議論が進んでも論点が交差せず，内容が似通ってしまう傾向がみられました．その結果，議論が発展せず，適切なタイミングで人間が介入して方向性を調整する必要が生じるという課題がありました．

このような背景を踏まえ，LLMエージェント発言多様化技術では「異なる視点から」「これまでに出ていない観点で」発言できることに加え，その発言が議論の流れに沿って自然に展開することも重視しています．これは，単に新たな視点の発言のみを追求すると話題から逸れてしまい，議論そのものが分かりにくくなってしまうからです．発言の多様さと議論に対する自然さの両立，つまり発言内容が「飛びすぎず」「古すぎず」というバランスをとることが，本技術の鍵となっています．

本技術では，**図2**に示す以下2点の方針に基づいて設計しています．

- あらかじめ用意された複数の発言指示（プロンプトテンプレート*4）に基づいて，1つのLLMエージェントから多様な候補発言を生成する（図2①）：発言候補の生成．
- 得られた複数の候補発言を，議論全体の整合性や新規性の観点から評価し，最適な発言1つを自動選択する（図2②）：発言候補の評価・評価に基づく発言選択．

以上の一連の処理により，人手を介さなくても，自然で具体性のある多様な対話の生成を実現することをめざしました．

■発言候補の生成：多様な発言指示テンプレートの活用

発言の多様性を確保するためには，LLMエージェントに「同じ議題に対して異なる角度で考えさせる」ことが重要です．本技術では，あらかじめ複数の発言指示テンプレートを用意し，これらを用いて1つのLLMエージェントから異なる方向性の発言候補を引き出します．

例えば，以下のようなテンプレートを設定します．

- 「ここまでの議論で合意できたポイントを整理したうえで意見を述べてください」
- 「今まで出た意見を整理したうえで深掘りする意見を出してください」
- 「他の視点から考えるとどうかを踏まえて具体案を出してください」

これらのテンプレートは，「方向性のずれ」や「深さの違い」を意図的につくり出すものであり，テンプレート自体が多様性を担保する仕掛けとなっています．

なお，テンプレートの種類や表現は議題や利用目的に応じて柔軟に変更可能であり，議論の幅を自在に調整することができます．

■発言候補の評価：自然さと多様性の両立

生成された複数の発言候補からどの発言

＊3 プロンプト：LLMに対する入力指示文．生成結果を誘導するための問いかけや指示文．
＊4 テンプレート：発言内容の方向性を定めるための定型的な指示文．多様性を確保する目的で複数使用します．
＊5 argmax：関数の出力が最大となる入力値を求める数学的操作．スコアの最大候補を選ぶ際に使用．

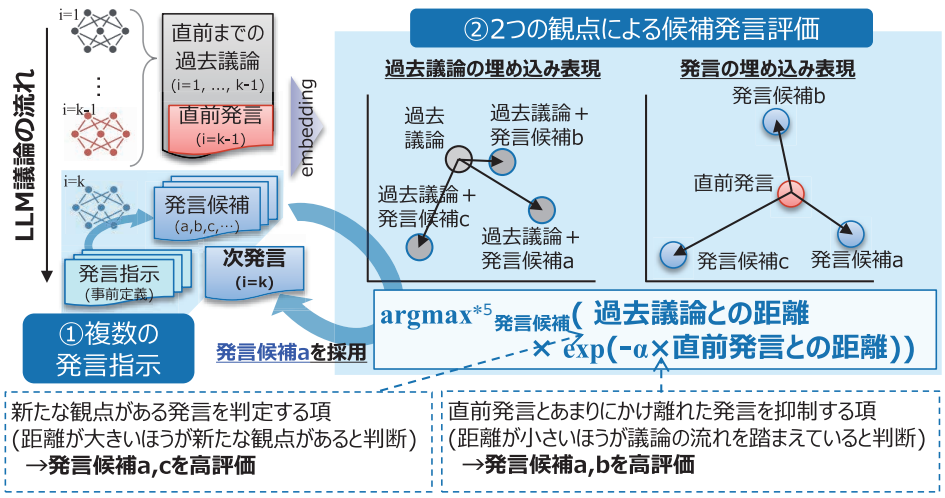

図2　LLMエージェント発言多様化技術

を選ぶべきかを判断するには，「これまでの議論とつながりがあるか」と「新たな視点が含まれているか」という 2 つの観点が重要になります．

この 2 つの観点を実現するために，本技術では発言を数値ベクトルとして表現（埋め込み：embedding）[*6] し，以下の 2 つの距離指標を用いてスコア化して評価します．

- 過去議論との距離：これまで出ていなかった新しい観点であるかを評価
- 直前発言との距離：議論の流れを踏まえた自然な発言かどうかを評価

例えば，過去議論と内容が大きく異なりつつ，直前の発言とは適度に関連がある候補発言は，「新規性がありつつ自然である」と判断され，高く評価されます．逆に，直前発言からの距離が極端に大きい発言候補や，過去とほぼ同じ内容の発言候補は低く評価されます．

■評価に基づく発言選択：自動で最も適切な発言を採用

発言候補に対して前述の評価を行った後，スコアがもっとも高い候補を 1 つ選択し，LLM エージェントの次の発言として採用します．この選択処理は完全に自動で行われ，ユーザによる確認や調整は不要です．

例えば，地域の高齢者支援を議題として LLM どうしに議論させ，これまでの発言から「高齢者に外出のきっかけを提供する」という方針が共有されており，直前の発言では「地域の体操イベントを増やすべき」との提案が出ていたとします．この状況で，以下の 3 つの発言が候補として生成されていたとします．

- 候補 A：「転倒防止のため，自宅でできるバランストレーニングを提案する」
- 候補 B：「近所での散歩を日課として推奨する取り組みをはじめる」
- 候補 C：「高齢者向けの栄養講座をオンラインで開催する」

このうち，候補 B は過去の発言と内容が近く，外出促進という文脈に沿っていますが，新規性には乏しいと判断されます．候補 C は全く異なる観点（栄養）を含み，新しさはあるものの，議論の流れからは逸脱

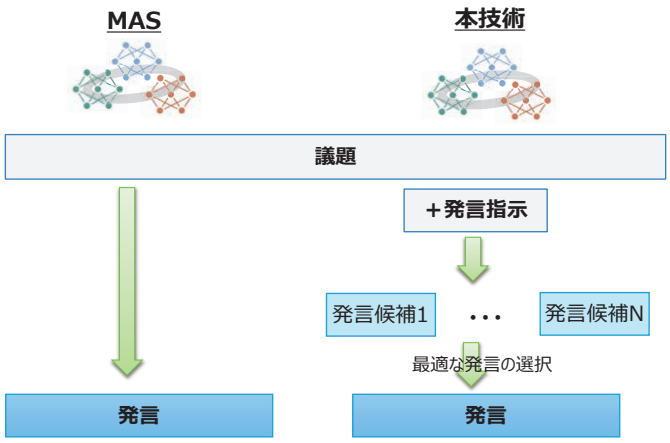

図 3　MAS と本技術の比較

していると判断されます．一方で，候補 A は直前の話題である「体操」に関連しながらも，「外出」からは離れた視点を提示しており，過去の議論とは違うが直前とはつながっているという構造を持ちます．そのため，新規性と自然さのバランスに優れた発言として高く評価され，最終的に採用されます．

このように，「新しい視点かつ議論の流れに自然な」発言が自動的に選ばれることで，単なる言い換えや繰り返しを避けつつ，意味的に発展性のある議論を実現していきます．また，一貫して同じ評価軸を用いて発言が選択されるため，議論全体に統一感が生まれ，理解のしやすさにもつながります．

評価実験：地方自治体における介護予防アイデア創出への適用

本技術を，地方自治体が直面する地域課題の 1 つである「介護予防」に関するアイデア創出の場面に適用して評価を行いました．具体的には，地方自治体における介護予防のステークホルダとなる 5 種類の LLM エージェント（例：作業療法士，民生委員，医師など）を設定し，それぞれの異なる立場から介護予防のアイデアを創出させて議論を行わせました．議論においては，本技術を用いて複数の発言生成と最適な発言の選択を実施しています．

技術の有効性を確認するため，次のような定量評価を実施しました．まず，従来の MAS 技術と本技術それぞれに対して，同じ議題と LLM エージェント設定の下で10回ずつ議論を実施し，各エージェントの発言を生成しました．各 LLM エージェントの発言には，アイデアについてのみでなく，そのアイデアを提示した理由やそれまでの議論に対する意見なども含まれます．そのため，各議論結果の「出力された発言そのもの」に加え，「出力された発言から LLM により抽出されたアイデアのみ」も評価対象として使用しています．

MAS 技術と本技術の違いを**図 3** に示します．LLM エージェントにより生成された発言（「出力された発言そのもの」と「出力された発言から LLM により抽出されたアイデアのみ」の両方を指す）から類似した発言を抑止できたかを評価するため，テキスト間の多様性評価手法である distinct-N[(3)] [*7]（生成されたテキスト中に含まれるユニークな N-gram [*8] の割合を測る指標，値が高いほうがいい）と，テキスト間の類似性評価手法である self-BLEU (Bilingual Evaluation Understudy) [(4)] [*9]

＊6 埋め込み：単語や文の意味をベクトルとして数値化し，機械的な比較や処理を可能にする技術．

＊7 distinct-N：文中に含まれるユニークな N-gram の割合．発言の多様性を評価する指標の1つ．評価においては N-gram の N=1に設定しました．

＊8 N-gram：連続する N 個の単語の組み合わせ．文章の多様性や特徴を定量化する際に用います．

（発言間の類似度をBLEU*10スコアに基づいて評価する指標，値が低いほうがいい）を用い，MAS技術との比較を行いました．具体的には，MAS技術と本技術の各結果を対戦させるかたちで合計100回比較し，各対戦においてdistinct-Nとself-BLEUを算出しています．

図4に評価結果を示します．出力された発言そのものを用いた場合，distinct-1はMAS技術の0.668に対して本技術は0.684に向上しました．同様に，self-BLEUもMAS技術の0.158に対して本技術は0.147に低下しました．また，出力された発言からLLMにより抽出されたアイデアのみを用いた場合も，distinct-1がMAS技術の0.633から0.644に向上し，self-BLEUがMAS技術の0.161から0.141に低下しました．これらの結果から，本技術により，議論の発言において使われる語彙が一層多様化し，MAS技術よりも同じような発言の繰り返しが減ることが確認されました．

本技術により論点が交差した自然な議論が行えているのか，という評価については，本技術で設定した地域課題の議題に関するワークショップに参加した関係者へのアンケート結果を使って行いました．詳細については「AIコンステレーション®のフィールドトライアルと展望」で述べますが，参加者にも違和感のない自然な議論が行えていることが確認されました．

このように，本技術は地域課題を検討するワークショップ等への適応において，実用的に活用できる可能性を持っています．これからは，より多様な自治体施策や合意形成支援への応用が期待されます．

今後に向けた展開と課題

本稿では，LLMエージェントによる議論において，発言の多様性と議論全体の自然な流れを考慮するための技術について紹介しました．発言指示テンプレートの活用と，自動的な発言候補評価・発言候補の選択という仕組みにより，人手による介入なしで，既存のMASを上回る結果が得られることを確認しました．

本技術は，LLMエージェントが多様な視点から解を創出する「AIコンステレーション®」の構想において，議論の中核を担う技術要素の1つです．今後は，具体的なユースケースへの適用に取り組むとともに，過去の議論や論点を適切に記憶・活用できる仕組みといった技術的課題に対する研究開発を推進していきます．

■ 出力された発言そのものを用いた場合

手法	distinct-N (N=1)		self-BLEU (N=1)	
MAS		0.668		0.158
本技術	Win	0.684	Win	0.147

■ 出力された発言からLLMにより抽出されたアイデアのみを用いた場合

手法	distinct-N (N=1)		self-BLEU (N=1)	
MAS		0.633		0.161
本技術	Win	0.644	Win	0.141

図4　MASに対する本技術との比較結果

＊9 self-BLEU：生成文どうしの類似度をBLEUで測定する手法．低いほど文の重複が少ないことを示します．評価においてはN-gramをN=1に設定しました．

＊10 BLEU：機械翻訳などの出力が人間による正解文とどれほど似ているかを，N-gramの一致度によって測定する自動評価指標のこと．

■参考文献

(1) Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch：“Improving factuality and reasoning in language models through multiagent debate,” Proc. of ICML 2024, pp.11733-11763, Vienna, Austria, July 2024.
(2) T. Liang, Z. He, Y. Du, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu：“Encouraging divergent thinking in large language models through multi-agent debate,” Proc. of EMNLP 2024, pp.17889-17904, Miami, U.S.A, Nov. 2024.
(3) J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan：“A diversity-promoting objective function for neural conversation models,” Proc. of NAACL-HLT 2016, pp.110-119, San Diego, U.S.A., June 2016.
(4) K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu：“BLEU: a method for automatic evaluation of machine translation,” Proc. of ACL 2002, pp. 311-318, Philadelphia, U.S.A., July 2002.

（上段左から）孫 晶鈺／井手 綾乃／吉田 司
（下段左から）渡邊 千紘／豊田 真智子／竹内 亨

本稿で紹介した技術は，人の議論をサポートすることを目的としています．私たちの技術により，議論の場に新たな視点や気付きを与え，課題解決に向けた具体的なアクションにつながるきっかけとなれば幸いです．

◆問い合わせ先

NTTコンピュータ＆データサイエンス研究所
革新的コンピューティングアーキテクチャ研究プロジェクト
コンピューティングアルゴリズム研究グループ