卒業論文

LLMマルチエージェント弁証法的推論による質問応答

指導教官　　村上 陽平 教授

立命館大学情報理工学部
先端社会デザインコース

増尾 柚希

2026年1月1日

LLM マルチエージェント弁証法的推論による質問応答

増尾 柚希

内容梗概

大規模言語モデル（LLM）を用いた質問応答システムは，さまざまなアプリケーション，システムで実用化が進んでいる．そして近年では,LLM に複数の価値観や立場を付与したマルチエージェント対話が提案され, 模擬的な議論や意思決定を生成する技術が登場している. 多文化共生社会においては, 質問応答は単なる情報検索にとどまらず, 異なる文化的・宗教的・倫理的背景をもつ人々の相互理解を促進する役割が求められている．しかしながら, 既存の LLM ベースのマルチエージェント対話は, 対立する立場をもとに勝敗を決めるディベート的議論が多く，最終的な結論も表層的な折衷案の提示に留まる．たとえば,「学校清掃を生徒が行うべきか職員が行うべきか」といったトピックにおいて, 現状の LLM は「日常清掃は生徒, 専門清掃は職員」といった妥協案を返す傾向があるが, これは価値観の本質的な対立を踏まえ，清掃を教育活動として再定義した生徒と職員の共同清掃のような止揚的な解決の回答生成は依然として実現できていない．そこで，本研究では，多様な価値観の対立を扱うトピックを対象に，LLM を用いたマルチエージェント対話によって新たな価値・制度・概念の再構成（止揚）を導く弁証法的な対話プロトコルを提案する．具体的には，kido, kurihara[1] らの記号論理に基づく既存の弁証法議論，および統合のアルゴリズムの手続きを参照し，その形式的プロセスを LLM マルチエージェントが自然言語で模倣できるように再構成した段階的対話モデルを設計する．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

議論の収束

質問応答のシステムに適用される推論のプロトコルは，その推論過程の終了を保証できなければばならない．こと弁証法的対話プロトコルにおいては，エージェント同士の議論を収束させることで，推論過程の終了を保証する．記号論理を用いて表現されたエージェントの知識ベースや，それらを用いた論理は手動で定義されるため，有限であり，同じ論理の再使用を許可しない限り，それらを用いて行う議論は必ず収束する．しかし，自然言語を用いて議論を行う LLM のエージェント同士の議論では，意味的に同義である論理の再使用を制限する必要があり，その制御はプロンプトによる指示だけでは難しく，議論の収束を

保証できない.

高次な止揚論証の生成

統合（止揚）は折衷ではなく価値の再構成による高次な概念の創出を意味するが，その成立条件及び構築アルゴリズムは自然言語では明確化されておらず，形式的定義だけでは意味的統合として妥当かを判断できない．また，kido, kurihara[1] の統合アルゴリズムでは，入力としてお互いに defeated された論証のみを入力としていたが，LLM では議論の収束が難しく，議論に決着をつけることがほとんどできないため，統合アルゴリズムをそのまま適用することは実用的に困難である．

本研究では，以上の課題を解決するために，プロトコル内で反復的にエージェントの議論が行われる「反論」フェーズにおいて，同一内容の議論をプロンプトで制限すると共に，反論回数の上限を設けた．上限に達した議論はその時点で終了し，その時点までの対話から対立軸を対話エージェントによって抽出する．これにより，お互いの議論に決着がついていなくても，対立軸の抽出から統合アルゴリズムを適用することができる．止揚論証の構築においては，kido, kurihara[1] らの統合アルゴリズムをプロンプトに落とし込み，入力された論証の特化（Specialization），汎化（Generalization）を行うことで，対立する立場の主張を 1 つに統合することができた．加えて本研究では，提案プロトコル組み込んだ弁証法的推論による質問応答システムを実装し，kido, kurihara[1] らの止揚論証の例との比較を行った．本研究の貢献は以下の通りである．

議論の収束

実装したシステムを用いて 100 個のトピックについて議論を行わせ，対話プロトコルは必ず収束することを確認した.

高次な止揚論証の生成

実装した質問応答システムを用いて kido, kurihara[1] らの対話例を実際に行い，最終的な止揚論証が意味的に一致することを確認した．その上で Anghel[2] らのルーブリック評価を行い,「独創性」,「一貫性」,「弁証法性」において，既存システムよりも高いスコアを示した.

LLM Multi-agent Dialectical Reasoning for Question Answering

Yuzuki MASUO

Abstract

LLM マルチエージェント弁証法的推論による質問応答

目次

第1章　はじめに

近年，大規模言語モデル（LLM）を用いた質問応答は，さまざまなアプリケーション，システムで実用化が進んでいる．同時に世界では，多文化共生がますます進んでおり，さまざまな価値観を持つ人々が共に暮らす社会ができている．質問応答には情報検索としての役割に加えて，そのような社会に適応するために，異なる価値観を持つ人々の相互理解を促進する，合意可能な回答を生成する役割も求められている．この要請に対し，複数の役割や価値観を付与したLLMエージェント同士の対話により結論を導くマルチエージェント議論の枠組みが提案されている．しかし既存手法の多くは，対立する立場をもとに勝敗を決めるディベート的議論，あるいは双方の主張を並列した表層的な折衷案に収束しやすい．その結果，対立を保持したまま共通項を抽出し，新たな価値・制度・概念の再構成として解決を導く止揚的な回答生成には至りにくいという課題がある．

そこで本研究では，記号論理で形式化された弁証法的推論の枠組みに基づき，複数のLLMエージェントが論証構築・相互反論・統合を段階的に実行する対話プロトコルを提案する．具体的には，Kido, Kurihara らの弁証法議論形式 [1] を参照し，記号論理で示された対話プロトコルと統合の手法を，2人のLLMエージェントを用いた弁証法的推論による質問応答システムに適用するために拡張する．プロトコルはフェーズは論証構築フェーズ，反論フェーズ，統合フェーズから構成され，論証構築とそれに対する反論は，2人の対話エージェントによって相互に行われる．エージェントの論証は結論（Consequent）と，それを導くための前提（Premises）によって成り立つ．そしてこのプロトコルを適用したシステムは，エージェントの論証の状態や，プロトコルの制御を行う．具体的に次のようにプロトコルは遷移する．

論証構築フェーズではまず片方のエージェント（AG1）が主論証（main argument）を構築し，反論フェーズでは相手のエージェント（AG2）が反論可能かを判断し，可能であれば相手の論証に対する反論（defeating argument）を構築する．反論を行うことができない場合，AG1 の main argument は正当化され（main argument justified），システムは終了する．AG2 が反論できる場合，AG2 の反論後，AG1 は同様に再反論の可否判断を行い，可能であれば相手の反論に対する再反論を構築する．AG1 が再反論を行うことができない場合，AG1

の main argument は敗北する（main argumnet defeated）．この反論・再反論を 1 回のエポックとし，システムで設定された最大反論回数に達するまで繰り返し行われる．最大反論回数に達しても main argument が justified・defeated どちらかの状態に遷移していない場合，その議論は保留状態（pending）に遷移する．AG1 の main argument の状態が defeated，もしくは議論が pending の状態の時，次に AG2 の論証構築フェーズに遷移し，同様の議論を行う．お互いの main argument に関して議論が行われると、kido，kurihara らの Specialization，Generalization を用いつつ，以下のように工夫を行うことで統合を行う．

本来 kido, kurihar の統合アルゴリズムは，お互いの main argument がどちらも defeated されている論証のペアに対して適用されるが，本研究では pending の議論に対しても統合アルゴリズムを適用することにした．理由は 2 つあり，1 つは LLM エージェントによる議論の性質上，議論の決着がつかず，pending の状態で統合フェーズに進んでしまうことである．2 つ目は，実世界での議論を考慮した時，議論の決着がつかないことは容易に考えられ，むしろその議論の中から対立点を抽出し，統合することが，本来の議論を通じた止揚を達成する弁証法的推論と考えられるからである．

では統合の手法だが，以上の状態の議論を kido, kurihara の統合アルゴリズムの入力にするために，AG1 は相互の議論収束後，これまでの議論を振り返り，それぞれの議論の対立軸を明らかにし，アルゴリズムの入力となるようにそれぞれの AG の論証を再構築する．この再構築は既存手法の Specialization にあたり，お互いのエージェントの立場でも受け入れられる条件を付与した形の論証を構築する．Generalization では再構築した論証のペアに対し，それらが共に共存する制度や，社会的な背景を提案する．これらの操作により，各エージェントの主張が議論を通してお互いの価値観を理解した上で妥協案を提案し，それらが共存する高次な解決策を生成する．

本研究の実装上の課題は，(1）議論の収束と，(2）高次な止揚論証の生成である．課題（1）に対しては，反論フェーズにおいて反論の最大回数を設定することによって、議論の収束を保証した．課題（2）に対しては，kido, kuriara ら [1] の Specialization，Generalization をプロンプトに落とし込むことで，記号論理で定義された既存手法と同様の止揚論証を生成できることを確認した．以降では，関連研究と課題設定を整理した上で，システムのアーキテクチャと提案プロトコルの詳細設計を述べ，評価と考察を通じて有効性と限界を明らかにし，最

後に結論と今後の課題をまとめる.

第２章　関連研究

2.1　記号論理に基づく弁証法的議論とその限界

対立する主張を形式的に扱う枠組みとして，弁証法議論に関する研究が行われてきた．

Sawamura，Umeda ら [2][3] は，Prakken, Sartor ら [4] の拡張論理プログラミングに基づき，弁証法議論の枠組み（Computational Dialectics） を提案している．この枠組みでは論証の構築と，相手の論証に対する攻撃（rebut, undercut）を段階的に行い，複数のエージェントが対話を通じて弁証法的に統合を行う議論フレームワークを形式定義している．ただし Sawamura，Umeda らは文献の中で，具体的な止揚論証の導出方法は神託（a sort of oracle）であると述べており、形式定義される止揚論証の意味的な実態や，それらの構築方法は未実装であった．

kido, kurihara ら [1] はこれらの課題を解決するため，形式定義された止揚論証を，Specialization 及び Generalization を用いて構築する手法を提案した．これにより，弁証法的な対話を通じて，止揚を行う形式的な操作が明らかになった．しかしながら，これらの文献における論証の表現は記号論理を用いて行われており，自然言語の解釈を行うことは困難である．特に，高度な止揚論証を記号論理で導出しても，それが実際自然言語でどう表され，どのような意味を持つのかを表現することは，未だ課題である．

2.2　大規模言語モデルによる弁証法的議論とその限界

そこで近年では，大規模言語モデル（LLM）を複数のエージェントに分担させ，異なる視点から推論を進めるマルチエージェント対話が注目されている．

Anghel ら [5] は，(1) 主張の生成，(2) 反論の生成，(3) 統合（synthesis）の生成という段階的な弁証法的推論フローを提案し，生成結果をルーブリック評価や解析によって検査する枠組みを示した．しかし，Anghel らの手法では反論が基本的に単発であり，多ターンの議論を通して対立点や前提の衝突を十分に掘り下げる設計にはなっていない．そのため，統合も「対立が整理された状態」を踏まえて構築するというより，限られた入力から一度で生成する形になりやすい．また，固定的な三段フローを超えて，より自然な多ターンの弁証法的交換へ拡張する必要があることも述べられており，弁証法的推論としては発展途

上である．

以上より，止揚（統合）を弁証法的に妥当な形で生成するには，統合に先立って多ターンの議論を通じて争点を明確化し，どの前提がどのように衝突しているかを構造化して取り扱う枠組みが必要である．次章ではこの課題に対し，議論フェーズと統合フェーズを明確に分離し，対立点の明確化を経て止揚論証へ接続する提案手法を述べる．

第3章　システムアーキテクチャ

3.1　概要

3.2　システム構成

第4章　弁証法プロトコル

4.1　概要

4.2　議論フェーズ

4.2.1　論証構築フェーズ

4.2.2　反論フェーズ

4.2.3　統合フェーズ

第5章　評価手法

5.1　評価方針

5.2　評価データ

5.3　評価手法

5.3.1　既存システムとの比較

5.3.2　ルーブリック評価

第6章　評価結果

6.1　評価結果

6.1.1　既存システムとの比較

6.1.2　ルーブリック評価結果

第7章　考察

7.1　考察

7.2　展望

第8章　おわりに

謝辞

本研究の遂行にあたり，ご指導・ご助言を賜った関係者の皆様に深く感謝する．

参考文献

[1] Sawamura, H. and Umeda, Y.: Computational Dialectics for Argument-based Agent Systems, *Proc. ICMAS 2000*, pp. 271–278 (2000).

[2] Kido, H. and Kurihara, M.: Computational Dialectics Based on Specialization and Generalization, *Proc. JSAI 2008*, pp. 228–241 (2008).

[3] Anghel, C. et al.: Multi-Model Dialectical Evaluation of LLM Reasoning Chains, *Informatics*, Vol. 12, No. 3, p. Art. 76 (2025).