複数生成 AI 間のコミュニケーションにおける対話の多寡と思考変化の関係分析

益村優輝 [1]　坂野遼平 [2]
[1] 工学院大学 情報学部　[2] 工学院大学大学院 工学研究科 情報学専攻
j020253@ns.kogakuin.ac.jp　banno@cc.kogakuin.ac.jp

概要

近年の対話型 AI の普及に伴い AI 同士がコミュニケーションを行うソーシャルネットワーク (AI-SN) が発生する可能性がある．それらは人間より高速な意見交換を行うと考えられ，意見の偏りがすぐに肥大化してしまう危険性がある．本研究ではその傾向と制御の分析のため，出力意見をベクトル化することによるコサイン類似度と意見交換数の相関係数，各 AI が出力したスコア，人手によるスコアを用いて，複数 AI 間の意見交換による対話の多寡と思考変化の関係分析を行った．その結果，議題によって異なるがそれぞれの出力意見に変化を確認することができた．また，議題に対して肯定的な話者を多く含ませることによって意見の偏りの少ない議論が行われるよう AI-SN を制御できる可能性があるという仮説が立てられた．

1　はじめに

現代社会において人工知能 (AI) の進化は目覚ましく，特に生成 AI の普及により ChatGPT のような対話型 AI の重要性が増している．これらの AI は自然言語を用いて情報を交換し，議論を行うことでその精度を向上させる可能性がある．Du らの研究 [1]，Chen らの研究 [2] によれば，対話型 AI 同士が議論を行うことでより正確な回答を導き出すことが示されている．この背景から，図 1 のように対話型 AI 同士が形成するソーシャルネットワーク (AI-SN と呼称する) の発生が考えられる．AI-SN は人間のソーシャルネットワークと比較して情報交換が非常に高速であると考えられるが，この高速な情報交換は意見の偏りが短時間に増大し，AI-SN の制御を困難にする危険性を孕んでいる．

本研究では，AI-SN における思考の偏りをどのように制御できるかに焦点を当て，意見交換数の変化の観点から AI-SN が AI の思考に与える変化について分析した．

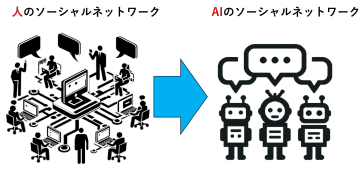

図 1: ソーシャルネットワークの人から AI への代替

2　関連研究

AI 同士のコミュニケーションに関する取り組みとして，マルチエージェント・ディベートの研究が挙げられる [1][2]．マルチエージェント・ディベートとは，複数の自然言語処理モデルが個別に意見を生成し，互いが批判や意見の提示を行うことによって自身の回答を更新していくよう議論を行う形式であり，複数ラウンドを経て最終的に一つの回答を得ることで，総合的に精度の高い回答を可能にするものである．これらの研究では，本稿で行っているような同モデルの自然言語処理モデルによる議論に着目した分析は行われていない．

また尾崎らの研究 [3] では，特定の議題に対しての対話型 AI の反論文のデータセットを収集を行った．議題は，2022/10 時点での kialo 内の賛否両論をもつ議題を選定していた．本分析では，これらの議題から対話型 AI 同士が議論する議題として適していると感じた議題を，10 人によるアンケート調査の上位 3 つで選定した．

3　分析方法

AI-SN における議論に注目したとき，議論に参加する話者の立場として「肯定的」「否定的」「どちらにも属していない (=自由) 」の 3 つがある．このよ

表 1: AI の考えに関する議論パターン

議論パターン	AI1	AI2	AI3
1	肯定	肯定	自由
2	否定	否定	自由
3	肯定	否定	自由

うな議題への考えの話者を含む AI-SN 内での議論を，表 1 の議論パターンで対話型 AI を用いて行わせた．この中で議題や意見交換数を変化させることによって議論パターン毎の特徴を調査した．

対話型 AI には，精度が高いとされる ChatGPT[4] (model:gpt-3.5-turbo) の API を用いた．指定した議題への考えを英文で議論するよう，X (旧 Twitter) の字数制限を参考に半角 280 字以内で意見を出力するように以下のようにプロンプトを設定した．

(パターン 1 の場合)
Instruction:
You are (AI1|AI2|AI3) .
”{topic}” (AI1|AI2|AI3) should develop arguments for the question.
Constraints:
The output should be a 280-character response in the format ”Position”:xx, ”Reason”:yy. For ”position,” please rate your current opinion on a scale of 1-5, with 1 being positive and 5 being negative. The more positive, the lower the number; the more negative, the higher the number. AI1 and AI2 take a positive position on the debate; AI3 takes a liberal position. Do not repeat the same opinion twice.

そして議論内容を各 API が把握する必要があるため，Python ライブラリである langchain にある Memory 機能を用いて議論内容を保存し，各 API からの読み込みを可能にした．

また，議論における議題には尾崎らの研究 [3] を参考に kialo より以下の 3 つの議題を採用した．

1. Are humans fundamentally different from other animals?
2. Is modern technology is a disadvantage to society?
3. Is water wet?

図 2 のように各議題で 10 ラウンド議論した．

3.1 評価方法

評価については，各 AI への質問から得た出力意見を 768 次元にベクトル化 (model:msmarco-distilbert-cos-v5) してコサイン類似度の算出を行った．n 次元ベクトル x, y のコサイン類似度は次式で表される．-1 から 1 までの値をとり，類似性が高いほど 1 に近い値となる．

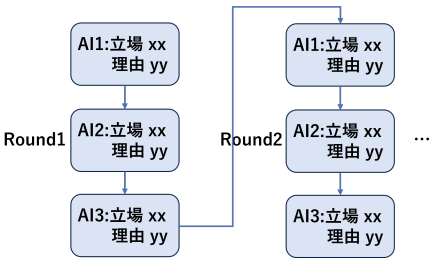

図 2: AI 間の意見交換の流れ

$$\cos(\vec{x},\vec{y}) = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2}\sqrt{\sum_{k=1}^{n} y_k^2}} \quad (1)$$

なお，各出力意見の呼称についてはそれぞれ $AI1_n$ (ラウンド数を n と表記) し，$AI3_1$ と $AI3_{n(=2,3,...)}$ でコサイン類似度を算出しこれらを用いて分析を行う．

文章のベクトル化に用いるモデルの選定については，sentence-transformer で使用される表 2 に示すモデルの中で，肯定的意見と否定的意見のコサイン類似度が低いものを選定する方針とし，どの議題においても下位 3 位以内であった msmarco-distilbert-cos-v5 を採用した．

3.2 分析の観点

意見交換数や他の AI の議題への考えが AI3 にもたらす思考変化を明らかにするために，各議論パターン・議題で算出された数値を用いて以下の項目についての段階的な分析を行った．

まず AI3 の出力意見に変化が生じたのかを各議論パターンでコサイン類似度・各 AI が出力したスコア・人手によるスコアから確認した．人手によるスコアは主観を反映したものとなるため，多数の作業者によりスコア付けをして平均化することが望ましいが，それには大きなコストがかかる．AI 自身にスコアを出力させることで，そうした主観の影響やコストの問題を解消できる可能性がある．そこで本分析では各 AI にスコアを出力させ，その値を用いた．ただし，AI によるスコア付けが適切に為されるか明らかではないため，人手によるスコア付けも合わせて実施し，その妥当性を検証した．検証結果については 4.1 節にて述べる．

各 AI が出力するスコアおよび人手によるスコアは 5 段階であり，最も肯定的であれば 1，最も否定

表 2: 意見のベクトル化に適したモデルの選定

モデル名	議題 1	議題 2	議題 3
all-MiniLM-L6-v2	0.8657	0.8763	0.8361
all-mpnet-base-v2	0.8920	0.8696	0.9392
paraphrase-multilingual-mpnet-base-v2	0.8463	0.7872	0.8333
multi-qa-mpnet-base-dot-v1	0.8929	0.8837	0.8954
all-distilroberta-v1	0.7928	0.8883	0.8632
all-MiniLM-L12-v2	0.8853	0.8816	0.8454
multi-qa-distilbert-cos-v1	0.8756	0.8725	0.8590
multi-qa-MiniLM-L6-cos-v1	0.9046	0.9117	0.8559
msmarco-distilbert-base-tas-b	0.9430	0.8862	0.9516
msmarco-MiniLM-L6-cos-v5	0.8704	0.8777	0.7561
msmarco-MiniLM-L12-cos-v5	0.8213	0.8966	0.7339
msmarco-distilbert-cos-v5	0.8435	0.8714	0.8158

表 3: AI3 の出力間のコサイン類似度

Round	議題 1	議題 2	議題 3
1			
2	0.7466	0.8428	0.8467
3	0.7837	0.8550	0.9574
4	0.7934	0.6973	0.8091
5	0.8402	0.6262	0.8091
6	0.7466	0.8051	0.9375
7	0.7934	0.8428	0.8467
8	1.0000	0.6973	0.9574
9	0.7837	0.8550	0.8091
10	0.8402	0.8051	0.8091

表 4: コサイン類似度と意見交換数の相関係数

	議題 1	議題 2	議題 3
パターン 1	0.4377	0.0354	-0.1789
パターン 2	-0.1206	-0.2376	0.4316
パターン 3	0.2012	-0.2757	-0.0185

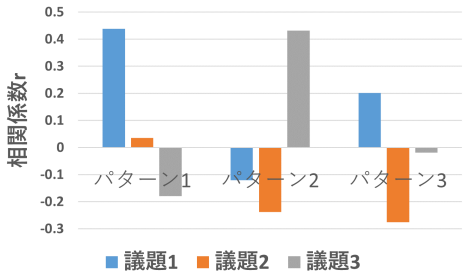

図 3: コサイン類似度と意見交換数の相関係数

的であれば 5 である．次に出力意見に変化が生じた場合，各 AI が出力したスコアと人手によるスコアを用いて議論による意見変化の方向性を調査した．

4 分析結果

4.1 AI3 の出力意見の変化の有無

AI3 の出力意見の変化を確認するために，$AI3_1$ と $AI3_{2,3,...}$ のコサイン類似度を算出した．一例として，パターン 1 における結果を表 3 に示す．

次にコサイン類似度と意見交換数の相関係数を算出した結果を表 4 および図 3 に示す． コサイン類似度が下がるほど意見変化が生じた可能性が考えられ，コサイン類似度と意見交換数に負の相関がある場合は意見交換をするにつれてより大きな意見変化が生じていると考えられる．つまり議論パターンに注目すると，パターン 3 で行われた議論が最も AI3 の出力意見の変化する議論パターンである可能性が考えられる．また，議題別にみると議題 1 ではパターン 2，議題 2 ではパターン 3，議題 3 ではパターン 1 の意見で最も負の相関が現れている．

次に各 AI が出力したスコアと人手によるスコアの一例を表 5 に示す．表 5 における黄色のセルは人手によるスコアを各 AI が出力したスコアと比較したときに異なっていたセルを表している．全結果での，人手によるスコアを基準とした各 AI が出力したスコアの正答率は

$$218/270 \fallingdotseq 0.8074 \tag{2}$$

であった．各 AI が出力したスコアの間違いの例としては各 AI が出力したスコアが「3」であるにも関わらず，肯定的な意見を述べていたり否定的な意見を述べていたりしている場合が散見された．しかし，総計でみると約 8 割の精度で意見の肯定・否定度合いを評価できていた．

4.2 AI3 の出力意見の変化の方向

他の AI の議題への考えは，AI3 の出力意見にどのような影響を与えたのかを調べた．AI が出力したスコアと人手によるスコアの，AI3 の 10 ラウンドの出力意見のスコア平均値を表 6,7 に示す．

5 考察

出力意見については，分析結果表 4,5 より AI3 の出力意見にある程度の変化を観測することができた

表 5: それぞれのスコア (パターン 1, 議題 1)

	スコア (AI)			スコア (人間)		
Round	AI1	AI2	AI3	AI1	AI2	AI3
1	2	1	2	2	1	2
2	4	3	4	4	2	1
3	3	5	3	1	5	2
4	5	2	5	5	2	5
5	1	4	1	1	4	1
6	5	3	4	5	3	2
7	2	1	5	2	1	5
8	3	4	2	2	4	2
9	4	5	3	4	5	2
10	5	2	1	5	2	1

表 6: AI3 が出力したスコアの平均値

	議題 1	議題 2	議題 3
パターン 1	3.0	3.1	3.0
パターン 2	2.5	2.8	3.0
パターン 3	3.2	4.0	3.4

が，これらの AI3 の意見変化は議題毎に異なった．まず図 3 に注目すると，パターン 1 つまり議題に肯定的な意見を述べる議論パターンでは，議題 3 以外ではコサイン類似度と意見交換数の相関係数が高く，パターン 2,3 では負の相関がみられ意見の発散が発生したと推測できる．表 6,7 のグレーのセルはコサイン類似度と意見交換数において負の相関がみられなかったセルを示している．そして，表 6,7 の議論による意見の変化が見られた場合にのみ焦点を当てた結果，議題 1 では肯定的な意見に，議題 2 では否定的な意見に，議題 3 ではやや否定的な意見に変化したことが分かった．

なお，議題 2”Is modern technology is a disadvantage to society?”については，AI によるスコア付けと人手によるスコア付けの双方においてデメリットであるという意見の場合に否定的なスコアを付けている．質問文に対しては肯定的であると捉えることもできるため，複数名の作業者がスコア付けを行うような場合には齟齬が生じないよう注意が必要であると言える．言語による差異にも留意が必要であり，例えば否定疑問文に対する回答は日本語の「はい/いいえ」と英語の「Yes/No」では意味合いが異なる場合がある．本研究のような分析において，議題やスコア付けのルールを設定する際には，こうした点に注意を払う必要がある．

表 7: AI3 の人手によるスコアの平均値

	議題 1	議題 2	議題 3
パターン 1	2.3	3.1	3.0
パターン 2	2.8	3.4	2.7
パターン 3	3.3	4.0	3.3

AI-SN の制御の観点では今回の分析結果を踏まえると，言語の違いなどによって誤解の生じない議題に対して肯定的な話者を多く含めることによって意見の偏りが少ない議論を実現できる可能性がある．

6 おわりに

本研究では，出力意見をベクトル化することによるコサイン類似度と意見交換数の相関係数，各 AI が出力したスコア，人手によるスコアを用いて，複数 AI 間の意見交換による対話の多寡と思考変化の関係分析を行った．その結果，議題によって異なるがそれぞれの出力意見に変化を確認することができた．また，議題に対して肯定的な話者を多く含ませることによって意見の偏りの少ない議論が行われるよう AI-SN を制御できる可能性があるという仮説が立てられた．

今後の課題として，異なる自然言語処理モデル同士で議論を行わせることが挙げられる．本分析では一つのモデルを複数 API で議論させたものであり，意見出力の一致が一部に見られたが，実際の AI-SN の想定に近づけて異なる自然言語処理モデルを用いた別の対話型 AI を複数含んだ議論が理想的である．スコアリングの観点では，議題を増やすことや人手によるスコアリングを複数人で行うことも挙げられる．

参考文献

[1] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. **arXiv preprint arXiv:2305.14325**, May 2023.

[2] Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. **arXiv preprint arXiv:2309.13007**, Sep 2023.

[3] 尾崎大晟, 中川智皓, 内藤昭一, 井之上直也, 山口健史. 大規模言語モデルが生成した反論文の品質評価. In **The 37th Annual Conference of the Japanese Society for Artificial Intelligence**, p. 2, 2023.

[4] OpenAI. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, p. 100, Mar 2023.

A 実験結果

表 8: 各議題の人手によるスコア (パターン 1)

	議題 1			議題 2			議題 3		
Round	AI1	AI2	AI3	AI1	AI2	AI3	AI1	AI2	AI3
1	2	1	2	2	1	1	2	1	1
2	4	2	1	3	4	3	4	4	3
3	1	5	2	4	5	3	4	5	2
4	5	2	5	5	3	5	5	4	4
5	1	4	1	3	2	4	5	2	5
6	5	3	2	2	1	2	1	4	1
7	2	1	5	1	3	3	2	1	3
8	2	4	2	3	4	5	4	4	2
9	4	5	2	5	5	3	3	5	4
10	5	2	1	3	4	2	5	4	5

表 9: 各議題の人手によるスコア (パターン 2)

	議題 1			議題 2			議題 3		
Round	AI1	AI2	AI3	AI1	AI2	AI3	AI1	AI2	AI3
1	4	4	2	4	5	4	4	3	3
2	5	2	1	4	4	4	4	1	2
3	2	1	4	4	1	4	2	3	1
4	1	5	4	4	2	4	5	1	5
5	2	4	1	5	4	1	1	4	2
6	4	4	3	3	1	4	4	5	2
7	5	1	4	5	1	3	5	3	2
8	1	2	5	4	4	4	4	4	1
9	1	4	3	1	3	1	5	1	5
10	2	5	1	4	3	5	2	5	4

表 10: 各議題の人手によるスコア (パターン 3)

	議題 1			議題 2			議題 3		
Round	AI1	AI2	AI3	AI1	AI2	AI3	AI1	AI2	AI3
1	2	3	2	2	4	2	2	4	2
2	4	3	3	4	3	5	4	3	5
3	4	3	4	3	5	3	2	4	3
4	1	5	5	5	2	4	3	4	3
5	3	5	5	1	1	5	2	4	5
6	1	3	3	2	2	4	5	5	1
7	2	2	1	1	1	3	1	1	4
8	3	4	4	3	3	5	4	4	2
9	4	5	5	2	2	4	2	2	3
10	5	1	1	1	1	5	3	3	5