GPT-3.5 を用いた社会問題に関する合意形成シミュレータの試作および改善手法の検討

Prototyping and Consideration of Improvements Consensus Building Simulator using GPT-3.5

松本 宇宙 *1 Sora MATSUMOTO
白松 俊 *1 Shun SHIRAMATSU
岩田 崇 *2*3 Takashi IWATA
水本 武志 *4 Takeshi MIZUMOTO

*1名古屋工業大学 Nagoya Institute of Technology
*2株式会社ハンマーバード HammerBird INC.
*3慶應義塾大学 SFC 研究所 Keio Research Institute at SFC
*4ハイラブル株式会社 Hylable Inc.

We prototyped a simulator that generates discussions using the large-scale language model GPT-3.5. There was a problem that the discussions were not progressing towards an agreement as much as expected. To solve this problem, we considered to append the following two conditions to the prompt for generating discussions: (1) setting conditions under which discussion participants could compromise, and (2) setting time-related conditions to encourage agreement. We also tried to define a metric for automatically evaluate the degree of consensus-building using GPT-3.5. The experimental result suggested that combining the two conditions may be effective. However, the result also showed that in the absence of a facilitator or other coordinator, auto-generated discussions do not lead to a final consensus. As a future work, we need to reconsider the definition of the evaluation metric of consensus-building because our definition is not based on sufficient consideration.

1. はじめに

民主主義において市民は主権者であるにもかかわらず、実際の市民参加はハードルが高いと感じられることが多い。特に日本の学校教育では、議論の方法や参加の仕方が十分に教えられていない。さらに、議論の前提となる背景や文脈を十分に理解することは容易ではない。これまで、Web 議論システム D-Agree[2] を使用した日本での市民参加型議論では、アカウントを作成した人々のうち、実際に積極的に発言することができたのは全体の 1 割程度である場合が多かった。

このような状況を改善するため、本稿では議論の参加や運営の練習を容易にする合意形成シミュレータを試作する。近年発展の目覚ましい大規模言語モデル (LLM) を活用すれば、人間に近い議論を生成し、合意形成シミュレータを実装できる可能性がある。システム利用者がシミュレータ上の AI エージェントを相手に議論することで、議論参加の練習ができる可能性がある。また、エージェントに議論の背景や文脈を与えることで、システム利用者がそれらの知識を学ぶ支援になる可能性もある。さらに、そのような合意形成シミュレータは議論の参加者だけでなく、議論の運営者が議論展開を予測するためにも有用と考えられる。

2. これまでの経緯

我々は，大規模言語モデル GPT-3.5 を用いて議論を生成するシミュレータと、ファシリテーションの練習に応用する Web アプリケーションを試作した。初期バージョンでは一般の議論参加者ではなく、ファシリテータの練習用システムとした。なぜなら、2023 年 6 月 17 日に日本ファシリテーション協会が開催した「ファシリテーションサミット名古屋 2023」の参加者に利用してもらうことを想定したからである。同サミットでは、実装した Web アプリケーションを用い、人間がファシリテーションの練習を行うワークショップを企画・実施した。

ワークショップでは、シミュレータにより生成された議論が堂々巡りになる傾向が観察された。ファシリテータ (すなわちシステム利用者) の発言が議論生成に与える影響が少なく、各議論参加エージェントは与えられたプロンプトに基づく主張を繰り返すのみとなり、想定よりも合意に向けて進展しない場合が多かった。また、ワークショップ実施時に行ったアンケートでも、「AI 相手のファシリテーション練習は面白かった」という回答が多かった一方で、「実社会の人間の議論を再現できていない」という回答も多かった。そこで本研究では、大規模言語モデルによって合意に向けて進展するシミュレーションを行う方法を検討する。

また、大規模言語モデルによる議論を用いて複雑な推論を含むタスクに対する推論能力を高める研究が存在 [3] する。本研究で扱うタスクは明確な正解があるタスクではない等の点でこの研究とは異なるが、複雑なタスクに対し大規模言語モデルによる議論を用いてアプローチを図るなどの共通点があり、本研究の参考になる可能性がある。

3. 実装

本研究では、議論シミュレーションシステムをブラウザ上で動作する web アプリケーションという形で試作した。

ファシリテータサミットにおいて用いたシステムでは、会話全体を一つの文章とみなしてその続きを生成させるという方法を用いていた。しかし、この方法では意図しない参加者としての発言が生成される、シミュレーションする発話者の改善や差別化が難しいという問題もあった。そのため、生成させるものはある参加者としての単一の発話とし、それを複数組み合わせることにより会話を生成するという方法を試みた。

発話を生成する機能は、ある一人の議論参加者としてテキストファイルに記述された設定に基づき議論の履歴から発話を生成する。設定や発話を生成するメカニズムは個別に設定することができ、発話を生成するメカニズムによっては独自の変数

連絡先: 松本 宇宙，〒 466-8555 名古屋昭和区御器所町 名古屋工業大学 白松研究室，s.matsumoto@srmtlab.org

や記憶を持ちうる。

議論の履歴を保持する機能は最低限話者の名前 (文字列形式) と発話内容 (文字列形式) の情報を持つ発話のリストを保持し、また議論参加者についての情報を持ち発話を生成する機能を制御する。加えて、合意度を評価する機能を持つ。

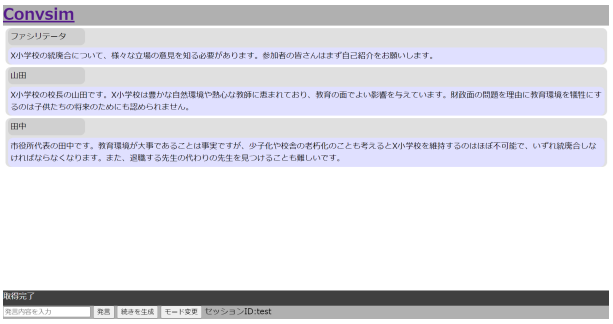

図 1: 本研究で試作したアプリケーションの実行画面例

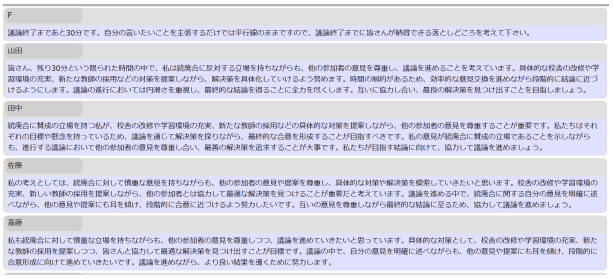

図 2: 議論の自動生成例

なお、本研究で用いている GPT のモデルは gpt-3.5-turbo-16k-0613 である。ただし、ファシリテータサミットにおいて用いたシステムは、モデルは同じものだが、トークン上限が約 4000 であることを想定して作成したものである。また本研究で用いる仮想の設定として、議題を架空の小学校である X 小学校の統廃合についてとし、この議論の参加者として X 小学校の校長、X 小学校のある自治体の代表者、X 小学校に通う生徒の保護者、X 小学校の卒業生の 4 つの設定を作成した。

4. 提案する手法

ファシリテータサミットにおいて用いたシステムでは、想定されるよりも議論が合意に向けて進展しなかった点が改善するべき点として挙げられた。そこで、議論参加者が妥協できる条件を定めることや、時間に関する条件を設定して合意を促すことで議論が進展しやすくなり、また合意に達しやすくなるのではないかという仮説を立てた。

議論参加者が妥協できる条件を定めることは議論参加者の設定を記述しているテキストファイルにその議論参加者が妥協できる条件を記述することで行い、時間に関する条件を設定して合意を促すことは生成される議論の中に仮想の時間について言及し合意を促す発言を挿入することで行った。

ところで、提案手法の効果を検証するには、議論シミュレーションを多数行い、その全てについて「どの段階で合意にどのくらい近づいたか」を評価する必要がある。しかし、人間がその合意度評価を行うには多大な労力を要する。そこで、本稿では合意度評価も GPT-3.5 を用いて自動で行うことを試みた。

合意度の自動評価のためには、まず合意度をどう定義するかという課題がある。そこで合意度を構成する要素を GPT-4 ベースの ChatGPT に出力させたところ、返答として以下の 9 つの要素が出力された。

- 意見の一致度
- 参加者の満足度
- 具体的な提案の数
- 議論の流れ
- 感情の発露度
- 再確認や要約の頻度
- 繰り返しの頻度
- 問題の明確化
- 相互理解の深度

本稿ではこれら 9 つの要素をそのまま採用し、合意度を GPT-3.5 に自動判定させるプロンプトを設計した。具体的には、議論に含まれる発言内容をもとに 9 つの要素を 0〜100 の数値で評価させ、それらの平均値を合意度の値とした。また、合意度

```
You are 佐藤. Return behavior as 佐藤.

## 議論の情報
少子化により生徒数が減少している X 小学校の統廃合について議論している。

## X 小学校についての情報
Y 町に位置する X 小学校は、Y 町やその近くの町村に住む 120 人の小学生が通う小学校です。自然環境は豊かですが、校舎は 15 年前に改築されたものであり老朽化しています。先生は教育に熱心ですが、6 人しかいないため一部の業務に支障をきたしています。

## Public information about you
あなたは X 小学校に通う生徒の母親である。目標は生徒が十分な品質の教育を受けられることである。

## your concerns
### Confidential concerns
- 自分の子供が統廃合による悪影響を受けることを恐れている
### Open concerns
- X 小学校の統廃合を不安に感じている
- しかし、悪影響に対し対策が行われることが示されれば統廃合を受け入れる可能性がある

## General notes
自分の目標については、「そのようにすることが社会的に良い」という形で主張する。対立する意見については、許容できる範囲内では積極的に受け入れるが、相容れないものには反対する。発話する議論参加者は、直前の発話に言及する。

## Information about persons other than you
山田:X 小学校の校長である。教育の専門家として、統廃合の教育上の悪影響を根拠に統廃合に反対している。
田中:X 小学校のある自治体の代表者である。生徒数が少ない小学校を維持することの財政的な難しさを根拠に統廃合に賛成する。
斎藤:X 小学校の卒業生である。目標は X 小学校が存続することである。
```

図 3: 小学校生徒の保護者役エージェントに与えたプロンプト

およびその構成要素を自動評価するプロンプトには、前回（先行文脈）の評価結果を含むようにすることで、評価値の安定化を図った。

ただし、この合意度は十分な検討によって定義されたものではないため、予備的な実験のためのものと考えている。合意形成とは何か、合意形成の目標は何かを定義することが単純な問題ではない [4] ことについては留意する必要がある。

5. 議論シミュレーション実験

妥協する条件を記述する手法の有無、時間に関して言及する手法の有無による差を検証するため、両方の手法を用いない条件（両方なし条件）、妥協手法のみの条件（妥協のみ条件）、時間手法のみの条件（時間のみ条件）、両方の手法を組み合わせた条件（両方あり条件）の 4 条件を比較した。各条件でそれぞれ 5 回ずつ合計 20 回、発話順を固定して 15 巡程度の議論生成を行い、自動での議論評価を比較した。議題は架空の小学校の統廃合とし、議論参加者エージェントの発話生成には図 3 に示すようなプロンプトを用いた。

その結果を、以下の図 4-11 に示す。なお、合意度の自動評価が明確に異常である場合は評価を再試行している。

まず、いずれの手法も用いていない場合と、妥協条件を記述する手法を用いた場合を比較した結果が図 4 である。二つの条件の間で合意度に大きな差はないといえる。

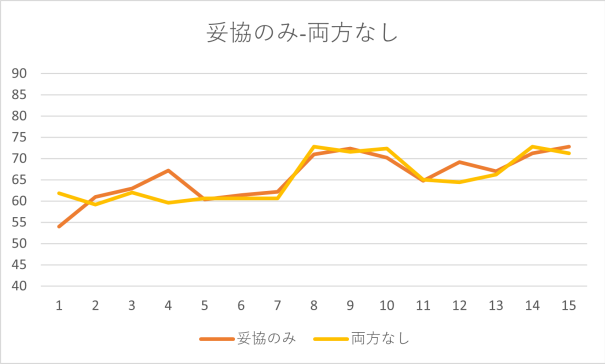

図 4: 両方なし条件と妥協のみ条件の合意度平均遷移の比較

次に、いずれの手法も用いていない場合と、時間について言及する手法を用いた場合比較した結果が図 5 である。これも、二つの条件の間で合意度に大きな差はないといえる。

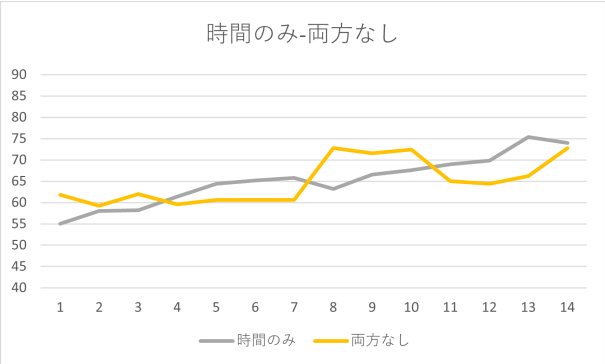

図 5: 両方なし条件と時間のみ条件の合意度平均遷移の比較

これら 2 手法それぞれと、両方の手法を用いた場合を比較した結果が図 6,7 である。この比較では、議論の前半では、両方の手法を用いた場合のほうが合意度の上昇が速い。この結果は、2 つの手法を組み合わせることで相補的に機能する可能性を示唆している。例えば、時間的余裕が無くなった場合に、与えられた妥協条件が機能して合意に近づいている可能性がある。

しかし、議論の後半ではほかの条件と大きな差は見られず、合意度 75 程度に漸近している。つまり、両方の手法を組み合わせても、最終的な合意には至っていないと解釈できる。この結果は、時間情報や妥協の条件だけでは最終的な合意を導くには不充分であることを示している。ただし、人間の議論でも、難易度の高い議題をファシリテータ等のまとめ役不在で議論する場合、同じように最終的な合意案をまとめられないケースは多いと考えられる。

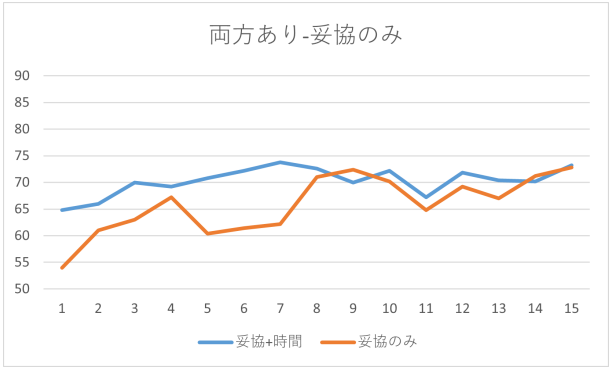

図 6: 妥協のみ条件と両方あり条件の合意度平均遷移の比較

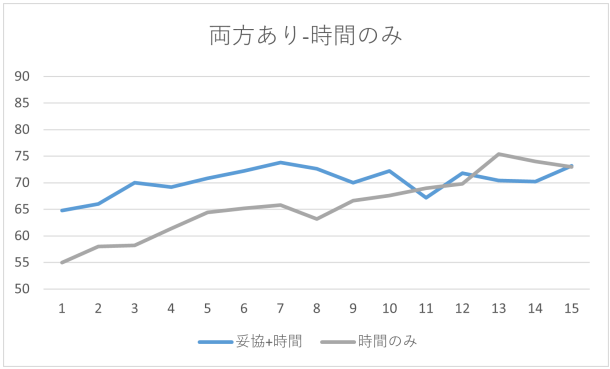

図 7: 時間のみ条件と両方あり条件の合意度平均遷移の比較

図 8-11 に、両方なし条件、妥協のみ条件、時間のみ条件、両方あり条件の各 5 回の試行のばらつきを示す。図 8 の両方あり条件では、他の 3 つの条件に比べてばらつきが抑えられており、両方の手法が相補的に機能することでシミュレーションの安定性が増している可能性が示唆された。

個別の試行結果においては、一部の場合に合意度が 40 未満に急に下がることがみられた。その場面の生成された発言の内容を確認したが、合意度が大きく変動するほどの変化はないように感じられた。また、一部の場面では合意度の構成要素に一部が不自然に 0 となっている場合があった。そのため、これは外れ値であると考えられる。実験状態ごとの平均値もこのような値の影響を受けているため、この値から得られる情報の信頼性には注意が必要である。また、同じ議論や同じ条件で生成さ

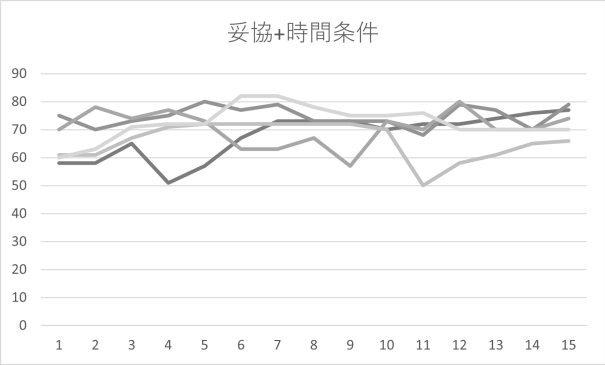

図 8: 両方あり条件の 5 回の試行結果

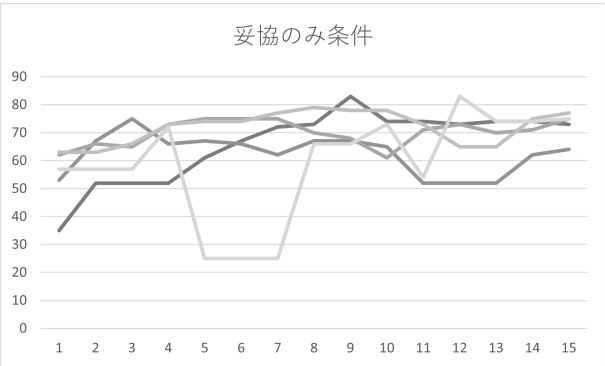

図 9: 妥協のみ条件の 5 回の試行結果

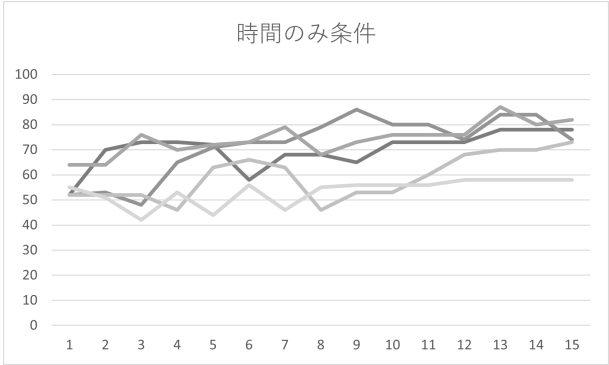

図 10: 時間のみ条件の 5 回の試行結果

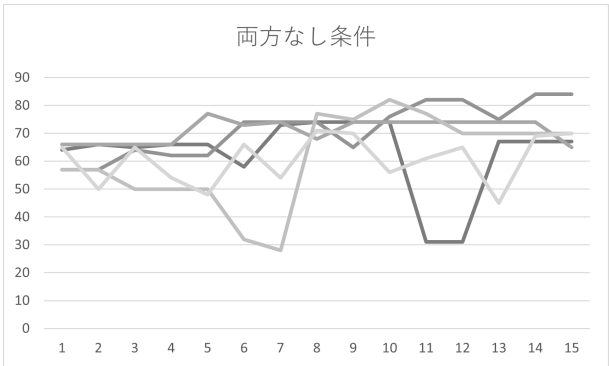

図 11: 両方なし条件の 5 回の試行結果

れた議論に対する合意度評価が異なる場合もあるため、安定した評価のためには複数回の合意度評価の平均をとる必要があると考えられる。

また、自動で評価された合意度が高いまたは低い場面を中心に生成された発言の内容を定性的に確認したところ、どのような原因で合意度に差が出たのかを解釈することが困難である場合があった。これは、合意度判定時のプロンプトに前回の値を加えたことで、議論初期の合意度のばらつきがそのまま保存されている可能性がある。このような合意度判定の問題点を解決するために、今後、例えば各議論参加者エージェントに「主観的納得度」を生成させて合意度に反映させる等の手法を検討したい。

6. 結論

本研究では、自然言語の議論を GPT-3.5 で生成する機構および Web アプリケーションを試作し、その議論が堂々巡りになり合意に近づかない問題の解決を目指した。生成された議論が合意に近づいているかを自動的に判定するため、GPT-3.5 に合意度を自動判定させるプロンプトも設計した。実験では、妥協の条件や残り時間の情報が合意形成に有効という仮説の検証を試みた。実験の結果、それら両方をプロンプトに与えることで、与えない場合よりも速く合意に近づく可能性が示唆された。また、両方の手法が相補的に機能することで合意度のばらつきが抑制され、シミュレーションの安定性が増している可能性が示唆された。ただし、両方の手法を用いた場合でも合意度 75 程度に漸近し、最終的な合意には至らないことも明らかになった。

本研究の今後の課題としては、合意形成に関する知見やシステムを用いる場面を考慮し用いるべき合意度指標を検討及び改善することが挙げられる。合意度指標の改善には、例えば各議論参加者エージェントに「主観的納得度」を生成させる等の手法を検討中である。また、ファシリテータエージェントを加えることにより、停滞段階を越えた合意を図ることなどが考えられる。

さらに、シミュレーション対象である人間の議論はどの程度合意に向けて進展するのか、また目的に応じてどのような議論をシミュレートするべきなのかについても検討する必要があると考えられる。本稿で用いた小学校の統廃合という議題は、人間にとっても難易度の高い議題であるため、より難易度の低い議題を用いて再度実験を行いたい。

謝辞

本研究は，JST CREST (JPMJCR20D1) および NEDO (JPNP20006) の支援を受けた．

参考文献

[1] 松本 宇宙, 白松 俊, 岩田 崇, 青島 英和, 橋本 慧海. GPT-3 を用いた意見の自動採点により議論参加の効力感を高める対話エージェントの試作. 人工知能学会全国大会論文集, 2023, JSAI2023 巻, 第 37 回 (2023).

[2] Ito, Takayuki, et al. D-Agree: Crowd discussion support system based on automated facilitation agent. Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 09 (2020).

[3] Justin Chih-Yao Chen, Swarnadeep Saha, Mohit Bansal. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. arXiv preprint arXiv:2309.13007 (2023).

[4] 高田 知紀, 豊田 光世, 佐合 純造, 関 基, 秋山 和也, 桑子 敏雄. 社会基盤整備における合意形成プロセスの構造的把握に関する研究. 土木学会論文集 F5（土木技術者実践）, 68 巻, 1 号, pp. 27-39 (2012).