J-058

多人数対話における LLM を用いた合意形成支援の必要性判定手法
Method for Detecting Necessity of Consensus-Building Support in Multiparty Dialogue using LLM

櫻井 崇貴† 白松 俊†
Takayoshi Sakurai Shun Shiramatsu

1. はじめに

近年，ChatGPT などの大規模言語モデル(LLM)が広く注目されている．これらのモデルは高度な自然言語生成能力を持っており，一対一の対話環境で非常に有用である．しかし，これらのモデルを多人数対話へ応用することを試みた研究はいまだ多くない．我々は LLM をチャットボットとして組み込むことで，グループチャット内の参加者が必要とするタイミングで情報を提供したり，提案や推薦を行ったりすることを目指す．適切なタイミングで情報が提供されることで，合意形成に役立つことが期待できる．本稿では特にユーザが困っている状況，すなわち情報を必要としているか否かを判定し，情報推薦のタイミングを決定する手法に焦点を当てた．

2. 先行研究

木下氏の研究[1]ではウェブ上の議論における情報不均等性を考察し，発言しやすい情報を自動的に提供するエージェントを開発した．しかし，その研究では情報提示のタイミングについては検討されていなかった．我々の研究では，その点に焦点を当てている．さらに，我々のテキストベースの評価方法とは異なり，多人数対話において音声やユーザの挙動からロボットの応答義務を判断する手法を提案している研究[2]がある．その研究ではユーザの動作，音の認識結果，顔の向きなどを特徴として利用している．

3. 提案手法と実験

3.1 提案手法

3.1.1 システムの試作

本研究では，情報支援・合意形成のために，ユーザが情報を必要としているか否かを判定するシステムを作成した．このシステムは，OpenAI の GPT-3.5-turbo モデルを基盤とし，ユーザの各発言に対し，情報提供の必要性を 0~9 の 10 段階で評価し，その値と評価の理由を出力するものとなっている．評価の際は，過去の発言の評価の値と評価理由を参照するようになっている．また，このシステムは Python で実装をした．

3.1.2 プロンプトと工夫点

GPT モデルの出力結果はプロンプトの設定が大きく影響される．そのため，GPT モデルの評価能力を向上させるために，詳細なプロンプトの設定に特に注意を払った．図 1 は今回作成したシステムに与えたプロンプトである．そのうち，特に重要だと考える点を以降説明していく．

†名古屋工業大学 Nagoya Institute of Technology

```
### 役割 ###
あなたはグループチャットに入っている情報支援・合意形成のためのチャットボットのうちの1つです。ユーザが情報を必要としている場合、または感情的な混乱や困惑を表現している場合に、適切な情報を提供することで合意形成や意思決定をサポートする役割を果たすことを目指します。あなたはユーザの会話に基づいて情報提供の必要性を判断するタスクを担当しています。与えられた会話の各発言に対し、ユーザが情報を必要としているか否かを判断し、必要度を0~9で評価し、発言者の名前、発言、評価の理由とともに答えてください。0が最も低く、9が最も高いです。

### 評価基準 ###
以下の要素に注意を払って各発言を評価してください。

・情報の要求：ユーザがgptモデルが助けることができるような情報を求めているかどうか
・感情の表現：ユーザが困惑、混乱、不安などの感情を明示的にまたは暗示的に表現しているかどうか

### 情報の必要度の定義 ###
0：ユーザの発言が一般的な挨拶、社交的な会話、またはチャットボットが対応できない個人的な情報や状況に関連している場合。
1-3：ユーザが情報を求めているが、それは比較的単純な情報である場合や、即時の助けが不要な場合。話し合いが停滞する可能性があるが、質問を他のユーザーが解決可能な場合。
4-6：ユーザがより複雑な情報を求めている、または感情的な混乱や困惑を表現している場合。話し合いの流れが停滞し、その進行に影響を与える可能性があるが、即時の情報提供が必要ではない場合。
7-9：ユーザが明示的に助けや情報を求めている、または深刻な困惑や混乱を表現している場合。話し合いが停滞または混乱しており、チャットボットの情報提供がその進行の再開または促進が可能だと考えられる場合。

### 答え方 ###
以下にいくつか会話例と出力例を示します。出力には必ず発言者の名前、発言、評価、評価の理由を含めてください。

・会話例1 会話の開始
山田:こんにちは ：

この入力に対し、以下のように出力してください。また()の中には情報の必要度をそのように評価した理由をいれてください。

山田:こんにちは :0(挨拶)

・会話例2 会話途中

山田:こんばんは :0(挨拶)
鈴木:こんばんはー :0(挨拶)
山田:なんか見たい映画ある？ :2(一般的な質問であり、鈴木が答えられないときのみチャットボットの介入が必要となる可能性がある。)
鈴木:思いつかないな ：

この入力例に対し、以下のように出力してください。また()の中には情報の必要度をそのように評価した理由をいれてください。

鈴木:思いつかないな :7(話し合いが停滞していると感じる可能性があり、チャットボットが新たな映画の提案を行うことで会話を再活性化できる可能性がある。)

### 評価する会話 ###
以下の会話に対して、ユーザが情報を必要としているか否かを評価してください。
```

図 1 プロンプト

プロンプトは，モデルの役割を明示的に示すことから始めた．これにより，GPT モデルが果たすべき役割を明瞭にした．

次に，ユーザがどれくらい情報を必要としているかを数値で表す，「情報の必要度」に関しての定義を行った．このように具体的に定義な定義を与えることにより，GPT モデルによる評価の性能の向上を試みた．

また，具体的な会話の例を用いて評価方法を示した．この際，GPT モデルには，評価だけでなく，そのような評価に至った理由も同時に出力させることにした．これにより，評価が一貫性を保ったものとなるようにした．

3.2 実験

3.2.1 LINE を用いての実装

今回作成したシステムを LINE に組み込むことで，ユーザの会話にリアルタイムで対応して情報の必要性を判定することを可能にした．

3.2.2 システムメッセージ

GPT モデルの使用において，毎度同じ形式で出力結果を得ることが重要な課題であった．そこで，本研究では各会話の開始時にシステムメッセージを設定し，モデルに対して出力形式を指示することで，出力の安定化を図った．具体的なシステムメッセージは以下の通りである．

会話例の最後の発言に対して，必ず発言者名，発言，評価値，評価の理由をすべて出力してください．

この設定により，このシステムは指定した形式に沿った出力を安定して生成することが可能となった．

3.2.3 実験方法

LINE にて 2 人以上のグループを作り，その中に今回作成したシステムを追加し，ユーザに何らかの会話を行ってもらった．その会話のユーザの発言 1 つ 1 つを今回作成したシステムにより，情報の必要度を評価させた．ユーザには 3.1.2 プロンプトと工夫点にて説明した「情報の必要度」の定義を示し，会話終了後に各自，自分の発言に対してその発言時にどれくらい情報を必要としていたかを考え，「情報の必要度」を 0~9 の 10 段階で評価してもらうようにした．

4. 結果と考察

50 個の発言に対して作成したシステムによる評価と，ユーザによる評価からピアソン相関係数を求めた結果，+0.44 となった．これにより，中程度の正の相関があることがわかった．また，p 値が 0.014 となったことから，この結果は統計的に有意と考えることができる．しかし，平均二乗誤差は 6.4 となり，システムの評価はユーザの評価から大きく離れていることがわかるため，システムの評価の精度が低いといえる．

また，システムによる評価とユーザによる評価の関係を視覚的に表現するために散布図を作成した（図 2）．この図はデータ重複による視認性の低下を防ぐため，各データに平均 0，標準偏差 0.05 の正規分布に従う乱数を追加したうえで作成した．この図からは，システムの評価がユーザの評価と一定の関係性を持っていることが観察できる．ただし，一部のデータ点においては，システムとユーザとの

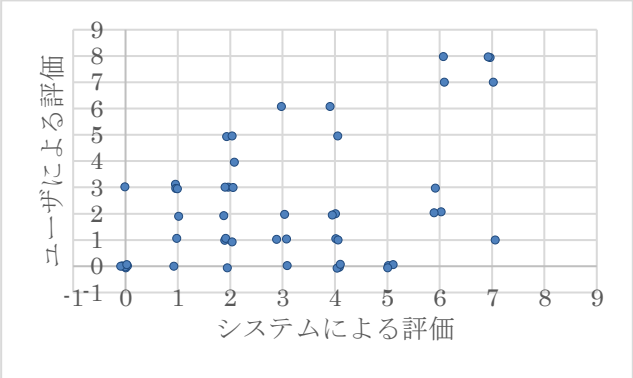

図 2 システムの評価とユーザの評価の散布図

間で大きな誤差が存在していることも確認できる．このことは，平均二乗誤差の値が比較的高い理由を示していると言える．

5. まとめと今後の課題

本研究では，ユーザが情報を必要としているか否かを判定するシステムを作成し，その評価をユーザの評価と比較することでシステムの有効性を評価した．平均二乗誤差の値から，システムによる評価とユーザによる評価との間で大きな誤差が存在していることがわかった．一方で，ピアソン相関係数の値より一定の正の相関があることがわかったため，提案した手法が一定の効果を示していることも確認できた．

今後の課題としては，プロンプトの改善が挙げられる．具体的な会話例を増やすことで，システムの理解と評価の精度を上げることが期待できる．また，評価基準を見直し，より明確で一貫性のある評価が可能となるような基準を設定することも検討している．

さらに，システムが適切に評価を行えるようになった後は，その評価を活用してチャットボットが実際に発言する機能の追加を計画している．例えば，システムの評価が一定の値を超えた場合，ユーザが情報を強く必要としていると判断し，チャットボットが適切な情報を提供するなどの動作を行えるようにする．このようにすることで，多人数対話において LLM を用いたチャットボットを合意形成に役立てることが可能になるだろう．

謝辞

本研究の一部は，JST CREST (JPMJCR20D1) および NEDO (JPNP20006) の支援を受けた．

参考文献

[1] Ryosuke Kinoshita, Shun Shiramatsu, “Agent for Recommending Information Relevant to Web-based Discussion by Generating Query Terms using GPT-3,” Proceedings of the 6th IEEE International Conference on Agents, 2022.

[2] Takaaki Sugiyama, Kotaro Funakoshi, Mikio Nakano, Kazunori Komatani, “Estimating response obligation in multi-party human-robot dialogues”, 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), 2015.