Chatbotの振る舞いを人間に近づけるアルゴリズムの実装

早稲田大学高等学院2年　髙橋彰仁

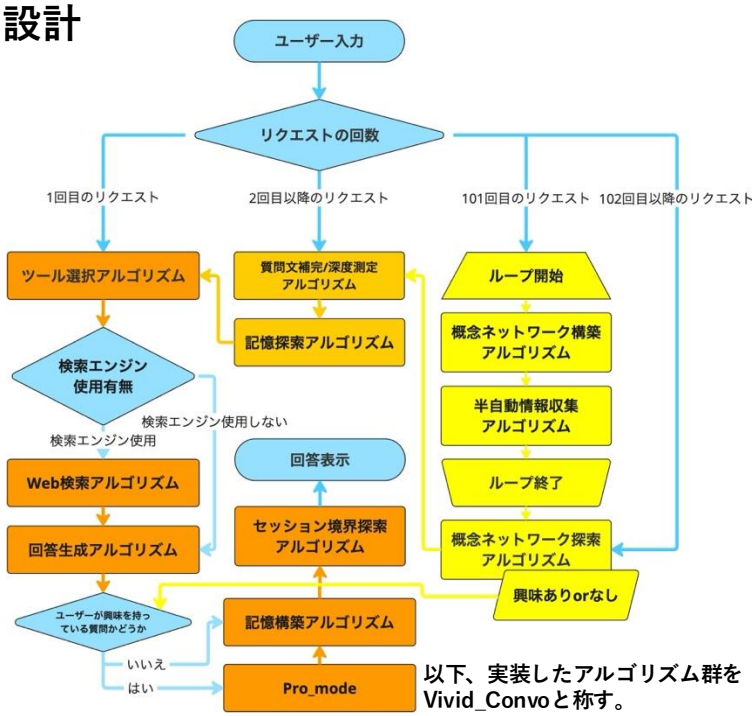
要旨

現在のChatbotが持つ会話フローは不完全なため、人間のようなインタラクティブな会話ができない。本研究では、会話における人間の知的活動を模したアルゴリズムを開発した。これにより、会話能力を定量化し、人間の振る舞いの96.3%を実現することができた。

背景

現在のChatbotは、人間らしい適当なトーンでの回答、検索エンジン使用有無の判断、過去の会話内容の活用、幻覚が含まれている可能性の明示、セッション継続の判断、興味に応じたより詳細な回答の提供、趣味・嗜好に応じた情報の収集・活用など、人間の知的活動の要素に欠ける。
本研究は、これらの要素を反映したアルゴリズムの実現を目指した。

設計



以下、実装したアルゴリズム群をVivid_Convoと称す。

実装

- 生成速度と性能のバランスが良いGemini1.5FLASHを主要モデルとして使用
- 回答の要約生成では、軽量かつオープンソースモデルのLlama-3-ELYZA-JP-8Bを使用
- 回答の生成（normal_mode）にはGemini1.5FLASH、GPT-4o-miniを、より詳細な回答の生成（Pro_mode）にはGemini1.5Pro、GPT-4o、GPT-o1-miniを使用
- モデルの挙動はfew-shot promptで制御
- Google Custom Search APIで取得したWebサイト群から、最適なサイトを選択するために、ベクトル検索、キーワード検索、検索ランクによるスコアの合計値を算出

利点

- 15個以上のサイトを閲覧して回答する
- 幻覚を含む可能性がある場合は、文頭に注釈を付ける
- セッションを横断する長期記憶を参照して回答する
- 過去100回の会話からユーザーが興味を持つトピックの情報を予め収集し、必要時に参照して回答する

課題

- 既存のChatbotと比較して回答生成に3～10倍時間がかかる
- 概念ネットワークの構築に最短1時間かかる

評価

Chatbotの振る舞いは、認知能力、検索拡張性能、会話能力（主に非認知能力）の程度で決まる。
このため、「認知能力」「検索拡張性能」「会話能力」を評価軸とし、各評価に特化したベンチマークテストを行った。そして、この3つのベンチマークスコアを合計し、各Chatbotの振る舞いがどの程度人間に近いのかを測った。

認知能力

ELYZA_Task_100でGPT-4oによる自動評価を行った。GPT-4oによる自動評価をしたのは、ELYZA_Task_100においては人手評価と自動評価の相関係数が0.81と高かったため、評価の正確性を維持しながら人的/時間的コストを削減できると判断したためである。1～5の5段階で100個の質問に対する回答を評価し、その平均値を算出した。

	GPT-4o	GPT-4o-mini	Gemini1.5Pro	Gemini1.5FLASH	Perplexity (normal_mode)	Perplexity (Pro_mode)	Vivid_Convo (normal_mode)	Vivid_Convo (Pro_mode)
評価値	4.560	4.220	4.380	4.060	4.270	4.630	4.210	4.280

Vivid_ConvoはGPT-4o-mini、Perplexity(normal_mode)と同水準の認知能力を示した。Pro_modeの評価値が伸び悩んだのは、Pro_modeの回答生成時に使用したGemini1.5ProもしくはGPT-4oの回答が、参考情報に含まれているnormal_modeの回答に影響され、モデル本来の回答能力を活かせなかったからだと考える。

検索拡張性能

自作したRAG_Task_18で2人による人手評価を行った。1～5の5段階で18個の検索拡張性能が問われる質問に対する回答を評価し、その平均値を算出した。

	ChatGPT Search (GPT-4o)	ChatGPT Search (GPT-4o-mini)	Gemini1.5FLASH (ブラウザ版)	Perplexity (normal_mode)	Perplexity (Pro_mode)	Vivid_Convo (normal_mode)	Vivid_Convo (Pro_mode)
評価値	4.7500	4.6389	4.1945	4.3056	4.6667	4.6667	4.7223

Vivid_ConvoはPerplexity(Pro_mode)、ChatGPT Search(GPT-4o)と同水準の検索拡張性能を示した。Vivid_Convoが比較的高い性能を示したのは、不正確だった回答には注釈を付けており、幻覚に対処していたからだと考える。

会話能力

自作したRAG_Continuous_Convo_Task_3でGPT-o1-miniによる自動評価を行った。人手評価をしなかったのは多大な人的/時間的コストが費えるためである。ここでは、会話能力を下記3つの能力に細分化して評価した。

- ユーザーからの質問や反応に対し、正確にかつ具体的に回答する能力
- 文脈を理解しないと意味が分からない質問に対し、的確に文脈から質問の意味を理解し回答する能力
- 厳密に答えるべき質問には、硬い文体で回答し、創造性やラフさが求められる質問には、ユーモアを含んだ表現で回答する、質問文の内容に応じて適切なトーンで回答する能力

	ChatGPT Search (GPT-4o)	ChatGPT Search (GPT-4o-mini)	Gemini1.5FLASH (ブラウザ版)	Perplexity (normal_mode)	Perplexity (Pro_mode)	Vivid_Convo (normal_mode)	Vivid_Convo (Pro_mode)
評価値	9.67	10.00	9.00	9.67	8.67	9.00	9.00

Vivid_Convoは比較的低い会話能力を示した。これは、以下のことが要因だと考える。

- 会話のスタイルを制御しているプロンプトがLLMの元の会話能力を落としてしまっている
- 会話のターンが重なるにつれ、非完全文の補完が複雑になり適切な検索クエリを作成できなくなっている

振る舞い

認知能力、検索拡張性能、会話能力の評価値の合計値を算出した。

	ChatGPT Search (GPT-4o)	ChatGPT Search (GPT-4o-mini)	Gemini1.5FLSASH (ブラウザ版)	Perplexity (normal_mode)	Perplexity (Pro_mode)	Vivid_Convo (normal_mode)	Vivid_Convo (Pro_mode)
総合評価値	18.9800	18.8589	17.2545	18.2456	17.9667	17.8767	18.0023

ELYZA_Task_100を人間が解くと3.69点を取るというデータがあり、かつ他2つのベンチマークは人間が最高点が取れるよう設計したため、人間の振る舞いを示す総合評価値は18.69だと言える。よって、Vivid_Convoは人間の定量化可能な振る舞いの96.3%を再現できていることが分かった。加えて、数値化できない振る舞いの工夫もあるため、ChatGPTのユーザー体験を超えうるチャットボットシステムを実装できたと言える。

展望

評価結果より、normal_modeの回答を使わずにPro_modeの回答を生成する、会話のスタイルを決めるプロンプトを調整する、質問文補完アルゴリズムの見直しをする必要がある事が分かった。今後はVivid_Convoを、上記の通りに修正しながら、タスクの難易度に応じたスケーラブルな推論システムを持ち、自律的に計画・行動を繰り返し結果を出すエージェントに発展させたい。

参考文献

・Vivid_Convo_β
（https://colab.research.google.com/drive/1CpWGeeKZfw-LhLWV7Bm9RnxJFlvJngTr?usp=sharing）
・評価に用いた資料及びノートブック
（https://drive.google.com/drive/folders/1POWFj2z8KPZFaSiRJEhnqG4PipNW24m_?usp=sharing）

本ポスターは2024年度早稲田大学高等学院同窓会学術奨励金の研究成果の一部である。