大規模言語モデルを使用したインタビューシステム

Research On Interview Robots Using Large-scale Language Models

三宅　芹奈 [1*]　河原　彩乃 [2]　下山　香音 [2]　奥岡　耕平 [3]
木本　充彦 [2]　今井　倫太 [1]
Serina Miyake[1]　Ayano Kawara[2]　Kaon Shimoyama[2]
Kohei Okuoka[3]　Mitsuhiko Kimoto[2]　Michita Imai[2]

[1] 慶應義塾大学　理工学部
[1] Faculty of Science and Technology, Keio University
[2] 慶應義塾大学院　開放環境科学専攻
[2] Faculty of Science and Technology, Keio University
[3] 日本大学　文理学部
[3] Nihon University College of Humanities and Sciences

Abstract: This paper focuses on a communication robot system aimed at facilitating interviews to gather information from humans, and it conducts research on a comprehensive method to elicit information. We propose the QCI-Robot, an interview robot system utilizing a large-scale language model. To validate the effectiveness of QCI-Robot, we conducted an evaluation experiment. The results demonstrate that QCI-Robot successfully generates questions that align with the context and topic, confirming the effectiveness of the proposed approach.

1　序論

人が他者から情報を引き出す方法としてインタビューが用いられる．インタビューは，構造化インタビュー，半構造化インタビュー，非構造化インタビューに分けられる [1]．構造化インタビューとは，インタビューの質問内容や順序があらかじめ決められているものである．半構造化インタビューとは，インタビューの主な質問事項は決めておくものの，順序は固定せず，個々の状況に応じて質問を変化させることができるものである．反対に, 非構造化インタビューとは，インタビューの方向性や内容がインタビュアーによって全面的に決定されるインタビューである．[2] インタビュアーが行う形式の多くは，事前に用意した質問項目をもとに質問を行う構造化インタビューまたは半構造化インタビューに当てはまる．

質問を列挙して行うインタビューの研究では，半構造化インタビュー，構造化インタビューを実現するために，質問生成（Question Gen-eration, QG）という技術を使用している．[4] では，ニュース記事のテキストを入力して，記事から抽出したキーワードから生成した質問項目を基にニューラルネットワークを使用して質問文を生成するインタビューシステムを提案している．さらに，大規模言語モデルを使用した会話の研究では，[5, 6] が特定の話題に焦点を当て深掘りする手法を提案している．しかし，ニューラルネットワークを使用して質問文を生成するインタビューの研究では，会話の文脈に沿った質問文の生成はできていなかった．また，大規模言語モデルを使用した会話の研究では，1つの話題を深掘りする質問が多く出力されるため，リストアップした項目に沿った質問は難しい．

本稿では，大規模言語モデルを活用し，事前に決定した質問項目に基づいて相手から情報を引き出すインタビューロボット QCI-Robot(Question Classifier Interview Robot) を提案する．QCI-Robot により，文脈に沿い，トピックに合った質問を行うことができる．QCI-Robot では，事前に用意した質問項目をトピックに関連する，関連しない項目に分類することで，トピックにあった質問をおこなう．また，大規模言語モデルを使用することで，文脈に沿った質問を生成することができる．

2　QCI-Robot

本稿で提案する QCI-Robot では，インタビュイーの音声を入力すると音声の内容に続くインタビュアーの質問文を出力する．QCI-Robot は，質問項目分類モ

*連絡先 ：慶應義塾大学理工学部今井研究室
〒 223-0061 神奈川県横浜市港北区日吉 3 丁目 14-1
E-mail:miyake@ailab.ics.keio.ac.jp

ジュール，質問選択モジュール，質問文生成モジュール，および質問判定モジュールで構成される．インタビュー開始前，質問項目分類モジュールで質問項目をトピックに関連するかで分類する．インタビュー開始後，質問選択モジュールで分類した質問項目のどちらから質問文を生成するかを判定し，文脈に合った質問選択を行う．次に，質問文生成モジュールで質問項目から質問文を生成する．関連する質問項目から質問生成を使用した場合，質問文をそのままインタビューに用いる．しかし，関連しない質問項目から質問文を生成した場合，質問判定モジュールで，生成した質問文がトピックに合っているかを評価する．生成された質問文がトピックに沿った質問である場合，関連しない質問項目からの質問文であっても，再利用してインタビューに用いる．すべての質問項目から質問文を生成すると会話を終了する．

QCI-Robot の構成図を図 1 に示す．QCI-Robot は大きく分けて質問項目分類モジュール，質問選択モジュール，質問文生成モジュール，質問判定モジュールの 4 つの機能で構成される．

2.1 質問項目分類モジュール

インタビュー中の文脈に沿った質問を可能にするため，事前に用意した多数の質問項目に対して，それぞれの質問項目が，事前に決定しているインタビューのトピックに関連するか，または関連しないかを会話開始前に分類する．関連しないと判断された項目は関連しない質問項目リストに格納する．また，関連すると判断された質問項目からランダムに 5 つの質問項目を選択し，関連選択質問項目リストに格納する．

2.2 質問選択モジュール

関連する質問項目，関連しない質問項目のどちらから質問文を生成するかを決定し，質問文を生成する質問項目を選択し，リストから削除する．ここで，関連選択質問項目リストに質問項目が入っていた場合は関連する質問から，リストに入っていない場合は関連しない質問から生成する．質問項目を選択時にリストから該当する項目を削除しているため，リストが空になる場合がある．関連選択質問項目リストに質問項目が入っている場合は，関連選択質問項目リストから質問項目を選択し，リストが空の場合は関連しない質問項目リストから質問項目を選択する．

2.3 質問文生成モジュール

インタビューの会話履歴全体と質問選択モジュールで選択した質問項目から，大規模言語モデルを用いて質問文を生成する．質問文生成モジュールでは，関連する，関連しないと判定された質問項目の両方を質問文生成の対象とする．関連する質問項目のみから質問を生成する場合，関連する質問項目として分類された質問項目は質問生成に用いられる可能性がある．しかし，関連しない質問項目として分類された質問項目は質問生成に用いられる可能性はない．つまり，質問項目がトピックに関連するかどうかの判定の際に誤って関連しないと判定された質問項目があったとしても，その質問項目から質問生成されることはない．そのため，関連する質問項目から生成した質問文は 5 つ出力し，関連しない質問項目から生成した質問文ではトピックに関連すると再判定された質問項目から生成した質問文のみ出力する．大規模言語モデルに入力したプロンプトでは大規模言語モデルに与えられた会話履歴に続く質問文を生成するように指示した．QCI-Robot ではこれまでの会話履歴を含めることで，相手の発言や文脈をより適切に理解し，質問文を適切に生成することができる．

2.4 質問判定モジュール

関連しない質問項目から生成した質問文がインタビューのトピックに関連するか否かを大規模言語モデルを用いて判定する．トピックに関連しない質問を防ぐため，トピックに関連すると判定された質問文のみ出力する．

3 実験

3.1 実験内容

QCI-Robot との会話がユーザに与える印象，QCI-Robot がトピックに合った質問文，文脈に沿った質問文を生成したかを評価するため，実験を行った（図 2）．なお，ロボットは Sota[7] を，マイクは YAMAHA 株式会社の YVC-300[8] を使用した．また，実験には，同じ研究室のメンバを含む慶應義塾大学理工学部の 8 人の学生が参加した (男性 6 人，女性 2 人，平均年齢 22.9 歳)．さらに，実験結果の客観的評価のアンケートでは，評価者として同じ研究室の学生 8 人が参加した (男性 5 人，女性が 3 人，平均年齢 23.0 歳)．

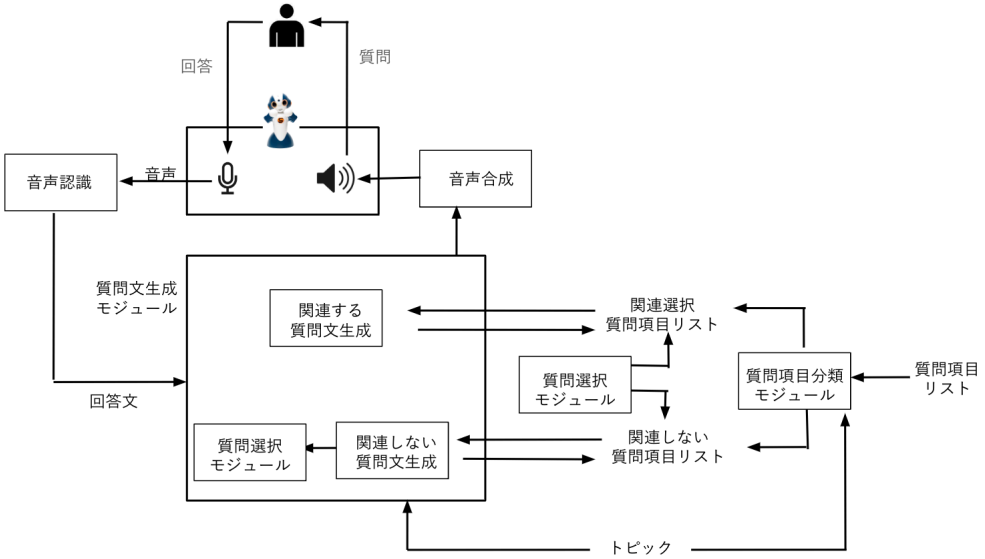

図 1: QCI-Robot のシステム構成

図 2: 実験の様子

3.2 条件

本実験では，被験者内計画で提案手法条件，関連質問条件，ベースライン条件の 3 つの条件について比較した．提案手法条件では，QCI-Robot を使用して関連する質問項目，関連しない質問項目から質問文を生成してインタビューする．関連質問条件では，関連する質問項目のみから質問文を生成してインタビューする．ベースライン条件では，質問項目分類を行わず，質問項目リストからランダムに質問項目を選択し，質問文を生成する．なお, ベースライン条件ではトピックに合った質問の割合に関する比較を行うため，提案手法条件と質問数が同じになるように設定した．

3.3 手続き

ロボットと実験参加者が 1 対 1 で対面形式でインタビューを行い，ロボットがインタビュアー，実験参加者がインタビュイーとなった．実験を行う前に実験参加者に 3 つの異なるトピックを用意してもらった．また，実験参加者はインタビュー終了後にロボットの印象に関する主観アンケートに回答した．一連の流れを，3 つの実験条件について繰り返した．実施した実験条件の順番はカウンターバランスをとった．

3.4 評価項目

インタビュー対話に関して主観的評価と客観的評価を行った．主観的評価では，実験参加者に対して主観アンケートで評価を行った．アンケート項目を表 1 に示す．実験参加者は，表 1 の主観アンケートを 7 段階のリッカート尺度で回答した．さらに，インタビュー実験後に実験参加者は 3 つのインタビューに対しての印象が良かった順に順番をつけた．

表 1: 主観アンケートの項目

ラベル	項目
Q1	ロボットは自分の話を聞いてくれていた
Q2	もう一度インタビューを受けたいと思う
Q3	ロボットの質問は適切だった
Q4	ロボットに対して話しやすかった
Q5	ロボットの質問に答えやすかった
Q6	ロボットは自分の話に興味を持ってくれた
Q7	ロボットの知識は豊富だった

客観的評価では，対面実験でのインタビュー履歴を基に，ロボットが行った質問のそれぞれが文脈に沿っているかを判断を評価者にしてもらった．評価者は「ロボットの発話内容全てがインタビュー会話の文脈に沿っているか」「ロボットの質問がインタビューのトピックに合っているか」の 2 つの質問に 7 段階のリーカット尺度で回答した．

3.5 結果の予測

主観的評価では，提案手法条件は関連質問条件よりもインタビュイーからトピックについての質問をより多くしており，提案手法条件，関連質問条件ではベースライン条件よりも文脈に沿った質問文を生成している．そのため予測 1 として，提案手法条件，関連質問条件，ベースライン条件の順にユーザからの印象が良いと考えた．

客観的評価では，提案手法条件は関連しない質問項目から質問文を生成しており，関連質問条件よりも質問数が多い．また，質問項目リストの中で，トピックに関連する質問項目の割合が半分以下になることが多いため，ランダムに質問項目リストから選択すると，トピックに関連する質問項目の割合が半分以下となると考えられる．そのため，ベースライン条件のトピックに関連する質問数は関連質問条件のトピックに関連する質問数よりも少ないと考えた．以上より，予測 2 として提案手法条件，関連質問条件，ベースライン条件の順によりトピックに関連する質問が多いと考えた．また，提案手法条件では関連質問条件より質問数が多いため，どちらの手法でも質問が 1 つずつトピックに関連しないと判断された場合，質問の全体数が多い関連手法条件の方がトピックに関連する割合は多くなると考えられる．一方，ベースライン条件では質問項目分類を行わないため，他の 2 条件に比べてトピックに関連しない質問項目からの質問が増えると考えられる．以上から，予測 3 として提案手法条件，関連質問条件，ベースライン条件の順にトピックに関連する質問の割合が多いと考えられる．さらに，トピックに関連しない質問をする際，文脈に沿った発話をすることが難しいと考えられることから，予測 4 として提案手法条件，関連質問条件，ベースライン条件の順により文脈に沿った質問文の割合が大きいと考えた．

3.6 結果

3.6.1 主観的評価

主観的評価において実験参加者が回答した主観アンケートの結果に対して，$\alpha = 0.05$ としたフリードマン検定を行った（図 3）．

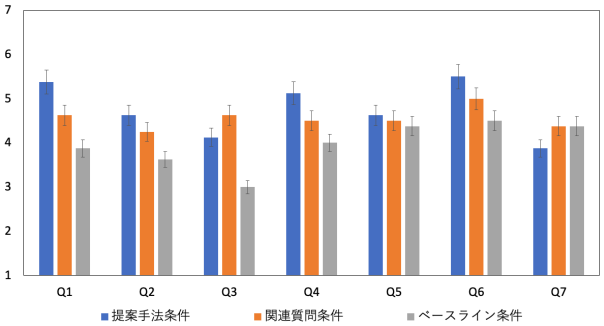

図 3: 主観アンケートの結果

図 3 より，3 つの条件において，Q1 から Q7 の各質問項目で有意差が示されなかった．(Q1:$F(2,14) = 3.44, p = 0.179$，Q2:$F(2,14) = 2.70, p = 0.259$，Q3:$F(2,14) = 5.786, p = 0.055$，Q4:$F(2,14) = 3.429$，$p = 0.180$，Q5:$F(2,14) = 0.071, p = 0.965$，Q6:$F(2,14) = 1.900, p = 0.387$，Q7:$F(2,14) = 0.583, p = 0.747$)

さらに，3 つの条件のインタビューを印象の良かった順に並び替え，それぞれ順位が 1 番，2 番，3 番となった回数について $\alpha = 0.05$ としたカイ二乗検定を行ったところ，提案手法条件，関連質問条件，ベースライン条件の順に 1 番となった回数が多く，有意差が示された ($\chi^2(2) = 18.75$，$p = 0.00088$)．

3.6.2 客観的評価

客観的評価において評価者が回答した客観アンケートに関して，文脈に沿っているか，トピックに合っているかの結果に対して，それぞれ $\alpha = 0.05$ としたフリードマン検定を行った（図 4，図 5）．

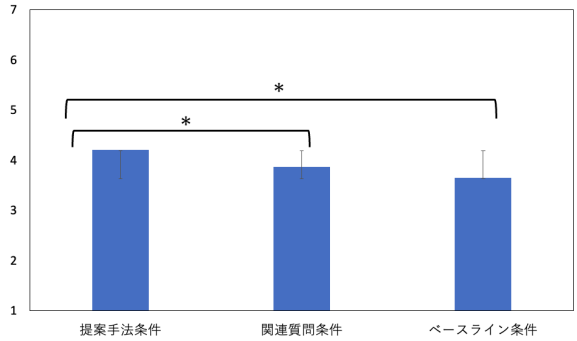

図 4: 文脈に沿っているかの結果

文脈に沿っているかについての分析結果によると，3 つの条件において，有意差が示された ($F(2, 62) = 10.859$，$p = 0.004$)．さらに，Bonferroni 法を用いて各条件の間についての検定を行った結果，提案手法条件とベースライン条件の間 ($p = 0.023$) 及び提案手法条件と関連質問条件の間 ($p = 0.035$) で有意差が示された．

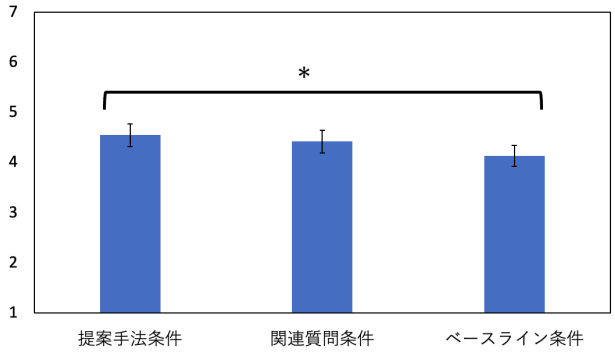

図 5: トピックに合っているかの結果

トピックに合っているかについての分析結果によると，3 つの条件において，有意差が示された ($F(2, 78) = 11.837$，$p = 0.003$)．さらに，Bonferroni 法を用いて各条件の間についての検定を行った結果，提案手法条件とベースライン条件の間 ($p = 0.01$) で有意差が示された．

さらに，客観アンケートでの結果から，トピックに合っていると判断された質問の割合の平均を比較した．ここでは，インタビュー履歴でのロボットの質問文の中で，客観アンケートのトピックに合ったかどうかの評価が 4 よりも大きい質問文をトピックに合っていると判断した．3 つの条件の客観アンケートのトピックに合ったかどうかの評価が 4 よりも大きい質問文の数について $\alpha = 0.05$ としたカイ二乗検定を行ったところ，提案手法条件，関連質問条件，ベースライン条件の間で優位差が示された ($\chi^2(2) = 12.75$，$p = 0.050$)．

4 考察

4.1 主観的評価

主観アンケートの各項目において，有意差が認められなかったものの，順位の検定結果より提案手法条件，関連質問条件，ベースライン条件の間でユーザのロボットに対する印象に関する有意差が認められたことから，予測 1 は支持された．

また，主観アンケートの項目に依らず，有意差が認められなかった原因として，3 つのトピックに対する答えやすさに差があったために結果に影響したと考えられる．トピックにたいする難易度を測ることは難しく，差が生じてしまった．そのため，3 条件について平等に評価することが難しくなったと考えられる．

4.2 客観的評価

4.2.1 トピックに合った質問

図 5 より提案手法条件とベースライン条件の間で有意差が認められたことから予測 2 は支持された．以上から，質問項目分類および，関連しない質問項目からの質問文生成が有効であったと判断できる．しかし図 5 より，トピックに合った質問の数においては 3 条件で差がないことから予測 3 は支持されなかった．

質問文の割合において有意差が認められなかった原因として，2 つ考えられる．1 つ目の原因は，ベースライン条件において質問項目をランダムに選択しているため，選び方によってはトピックに関連する質問項目の割合が多くなったことである．2 つ目の原因は，質問項目から質問文を生成する際に前置きなく質問を行うため，関連する質問項目から生成した質問文でもトピックに関係ないように捉えられることがあることである．

4.2.2 文脈に沿った質問

図 4 より，提案手法条件と関連質問条件及び提案手法条件とベースライン条件において有意差が認められたことが確認できる．以上のことから予測 4 が支持され，質問文の生成時に関連する質問項目からの質問生成後に関連しない質問項目からの質問文の生成を行うことが有効であると判断できる．

5 結論

本稿ではインタビューにおいて，事前に用意した質問項目について文脈に沿いトピックに合った質問を行うことを目的として，大規模言語モデルを活用し質問

項目に沿ったインタビューを行うロボット QCI-Robot を提案した．また，評価実験を行い，QCI-Robot の有効性が確認された．

謝辞

本研究の一部は，JST，CREST，JPMJCR19A1，JSPS 科研費 JP23K16980 の支援を受けたものである．

参考文献

[1] ASMARQ: 半構造化インタビューとは？メリットや実施時の注意点を解説　定性調査, *https*：*//www.asmarq.co.jp/column/column* − *cat/glossary/qualitative*3*/semi* − *structured*$_i$*nterview/*(2022)

[2] Jacqueline Pei, Jenelle M. Job, Cheryl Poth, Erin Atkinson:Qualitative research and evaluation methods (3rd ed.),*Evaluation Journal of Australasia*,pp.60-61(2002)

[3] Herbert J.Rubin , Irene S.Rubin: Qualitative Interviewing (2nd ed.): The Art of Hearing Data,*SAGE Publications*,pp.113-117(2005)

[4] Duan Nan ,Tang Duyu, Chen Peng,Zhou Ming: Question Generation for Question Answering,*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*,pp.866-874(2017)

[5] Tom B. Brown,Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei: Language Models are Few-Shot Learners,*Computation and Language*(2020)

[6] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, Ji-Rong Wen: A Survey of Large Language Models,*Computation and Language*(2023)

[7] ヴイストン株式会社:*https*：*//sota.vstone.co.jp/home/*(2024.1.12 閲覧)

[8] YAMAHA 株式会社:*https*：*//sound* − *solution.yamaha.com/products/uc/yvc* − 300*/index*(2024.1.17 閲覧)