

語り手の発話の言い換えにより語りに傾聴を示す応答の生成

茂木光志¹ 伊藤滉一郎² 村田匡輝³ 松原茂樹^{2,4}

¹名古屋大学情報学部 ²名古屋大学大学院情報学研究科

³豊田工業高等専門学校情報工学科 ⁴名古屋大学情報基盤センター

motegi.koshi.h9@s.mail.nagoya-u.ac.jp ito.koichiro.z5@f.mail.nagoya-u.ac.jp

murata@toyota-ct.ac.jp matsubara.shigeki.z8@f.mail.nagoya-u.ac.jp

概要

語る機会を創出するために、会話エージェントが語り手の聴き手を担うことが期待されている。これらが聴き手として認められるためには、語り手に対して傾聴態度を示す発話（傾聴応答）を生成することが有効である。傾聴応答の1つに、語り手の発話を言い換える応答（言い換え応答）がある。言い換え応答を適切に生成できれば、語りを理解していることを示すことに寄与する。本論文では、大規模言語モデルによる言い換え応答の生成可能性を検証する。生成実験を行い、その性能を評価した。

1 はじめに

語ることは人間の基本的な欲求である。語るためには、聴き手の存在が不可欠であるものの、日本では独居世帯の増加 [1] などによって、聴き手不在の場面が増加しており、語る機会が失われつつある。この解決策として、コミュニケーションロボットやスマートスピーカーなどの会話エージェントが語り手の聴き手を担うことが考えられる。これらが聴き手として認められるには、語りを傾聴していることを語り手に伝達する必要がある。そのための明示的な手段の1つは、発話の生成によって語りに応答することである。以降では、傾聴を示す目的で語りに応答する発話を**傾聴応答**と呼ぶ。

傾聴応答は、機能や形態によって、いくつかの種類に分類できる。傾聴応答の代表例は相槌であり、その生成手法の提案が行われてきた [2, 3, 4, 5]。ただし、語りの内容を必ずしも理解していなくても、相槌を表出することはできる。一方で、語りの内容を理解していなければ適切に表出できない応答も存在する。その1つが、語り手の発話に含まれる語句を言い換える応答、すなわち、**言い換え応答**である。言い換え応答を適切に表出できれば、語りを理解し

表1 語りと傾聴応答の例

語り	傾聴応答	種類
趣味の1つは	はい	相槌
園芸植物を	園芸	繰り返し
栽培することです	いいですね	評価

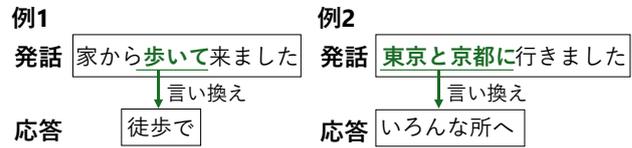


図1 言い換え応答の例

ていることを語り手に示すことができ、語りの傾聴をより効果的に伝えることができると考えられる。

そこで本論文では、言い換え応答の自動生成の実現に向けて、言い換え応答の生成可能性を検証する。語りに対する傾聴応答が収録されているデータを用いて、大規模言語モデルによる言い換え応答の生成実験を行った。生成された応答を人間が表出した応答と比較することにより、その性能を評価するとともに、言い換え応答の自動生成の実現に向けた課題を考察する。実験の結果、大規模言語モデルに基づく生成手法は、一定の水準であることを確認した。

本論文の構成は以下の通りである。2章では、傾聴応答と言い換え応答を説明し、関連研究について述べる。続く3章では、言い換え応答の生成手法を説明する。4章では、言い換え応答の生成実験について報告する。最後に、5章で本論文をまとめる。

2 語りの傾聴と言い換え応答

本章では、傾聴応答の1つである言い換え応答とその関連研究について述べる。

2.1 傾聴応答

傾聴応答とは、傾聴を示す目的で語りに応答する発話であり、適切に表出できれば、語り手の語る意

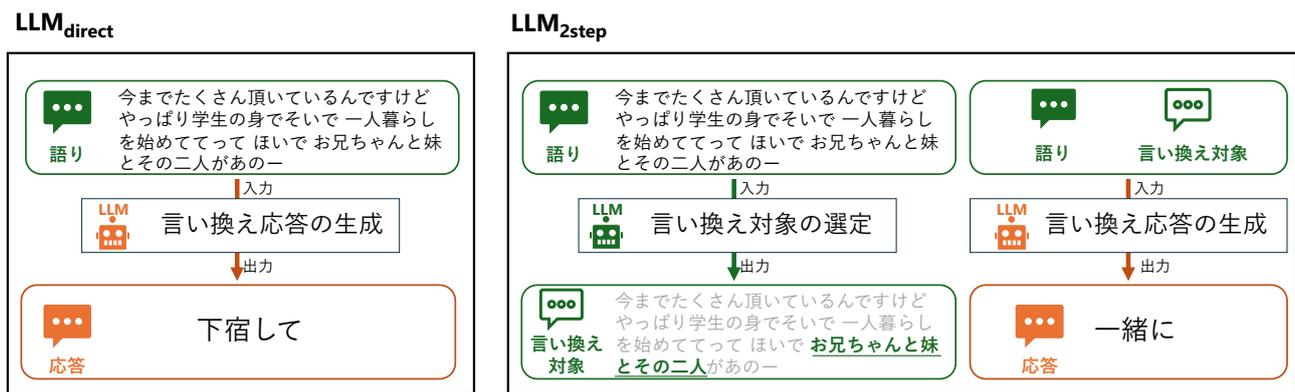


図2 言い換え応答生成手法の概略

欲を高める効果が期待できる。表1に、語りと傾聴応答の例を示す。傾聴応答は、機能や形態によって、いくつかの種類に分類することができる。表1に例示するように、相槌、繰り返し、評価などの応答が存在する。

2.2 言い換え応答

傾聴応答の1つに、語り手の発話に含まれる語句を言い換える応答、すなわち、言い換え応答が存在する。図1に、語り手の発話に対する言い換え応答の例を2つ示す。例1では、発話に含まれる「歩いて」を「徒歩で」に言い換えている。例2では、発話に含まれる「東京と京都に」を「いろんな所へ」に言い換えている。語り手の発話に応じた適切な言い換えを行えば、語り手の発話の理解を相手に示すことができ、語りを傾聴していることを効果的に伝えられるものと考えられる。

2.3 関連研究

傾聴応答にはいくつか種類があり、その代表例が相槌である。相槌については、その生成タイミングの検出手法が提案されてきた。これまでに、ルールベースによる手法[2]、n-gramモデルによる手法[3]、SVMによる手法[4]、事前学習済みの言語モデルによる手法[5]などが提案されている。

相槌以外の特定の傾聴応答に着目した研究も存在する。例えば、語り手の発話に含まれる語句を繰り返す形式の応答、すなわち、繰り返し応答に関して、その表現を生成するTransformer[6]ベースの手法が提案されている[7]。そのほかにも、語り手の発話に対する評価を示す応答、すなわち、評価応答に関して、その応答タイミングの検出や応答表現の選定を行う手法が提案されている[8,9]。また、語り手の発話に同意しないことを示す応答、すなわ

ち、不同意応答に関しても、その応答タイミングの検出や応答表現の生成が試みられている[10]。

このように、傾聴応答の自動生成に向けた研究に取り組んだ研究がいくつか行われているものの、言い換え応答については、その生成に関する検討が十分になされていない。

3 言い換え応答の生成手法

言い換え応答を適切に生成するには、応答タイミングの検出と、応答表現の生成を適切に行える必要がある。これら2つのうち、本研究では、応答表現の生成に着目し、言い換え応答タイミングを既知として、言い換え応答タイミングにおける言い換え応答の表現の生成可能性を検証する。応答表現の生成手法を実装し、その性能を後述する実験によって評価する¹⁾。

言い換えについて、これまでに、ルールベースによる手法[11]、LSTM[12]に基づく手法[13]、事前学習済みの言語モデルに基づく手法[14]など多数提案されている。本研究では、大規模言語モデルに基づく以下の2つの手法を採用した。図2に、その概略を示す。

- LLM_{direct} : 語り手の発話を大規模言語モデルに入力し、言い換え応答を直接生成する手法。
- LLM_{2step}: 言い換え応答を2段階で生成する手法。まず、大規模言語モデルを用いて、語り手の発話に含まれる言い換え対象の語句を選定する。次に、語り手の発話に加え、選定した言い換え対象の語句を入力として、言い換え応答を生成する。

言い換え応答を生成するためには、言い換え対象と

1) 以降では、特に断りがない限り、言い換え応答の生成とは、言い換え応答表現の生成を指すものとする。

表2 システムプロンプト (共通)

role	content
system	あなたは役に立つアシスタントです。 あなたは発話に対して傾聴応答を行うことができます。傾聴応答は発話に対する傾聴態度を示すために行われる発話を指します。 言い換え応答は、発話の一部を別の言葉を用いて表現する傾聴応答です。

表3 言い換え応答を生成するプロンプト (LLM_{direct})

role	content
user	次の発話に対する言い換え応答を考えてください。 発話：{語り手の発話}

表4 言い換え対象を選定するプロンプト (LLM_{2step})

role	content
user	以下に話し手の発話を与えます。 語り：{語り手の発話} 与えた語りの後に言い換え応答を表出する場合に、適切な言い換え対象となる語りの一部を抜き出してください。

表5 言い換え応答を生成するプロンプト (LLM_{2step})

role	content
user	次の発話に対する言い換え応答を考えてください。指定した箇所に着目して、言い換えてください。 発話：{語り手の発話} (指定範囲：{言い換え対象の語句})

なる語句の選定を選定することと、選定された語句を言い換えることが必要である。言い換え応答の生成を、それらの2つのタスクに分割することにより、モデルが各タスクに注力できるようになる効果が期待できる。

4 実験

言い換え応答の生成可能性を検証するために、傾聴応答の収録データを用いて、生成実験を実施した。

4.1 実験データ

実験データには、傾聴応答コーパス [15] を使用した。このコーパスには、語りに対する傾聴応答が収録されている。語りのデータには、高齢者のナラティブコーパス JELICo [16] が用いられている。また、語りと傾聴応答には発話時間が付与されている。傾聴応答コーパス内の応答には、その種類がラベル付けされている。ラベルの候補は16通り存在し、その中には言い換え応答も含まれる。傾聴応答コーパスには、1,170個の言い換え応答が含まれており、学習データ、開発データ、テストデータに分割されている。本実験では、開発データの言い換え応答258個を評価用データとして利用した。

本実験では、語りの発話単位として、述語を中心としたまとまりである節を採用した。節境界解析ツール CBAP [17] を用いて、語りデータを節に分割した。言い換え応答の発話開始時刻と、語りの節の発話終了時刻に基づいて、応答と節を対応付けた。言い換え応答の生成時には、応答に対応付いた節を含む、直前の5つの節を入力とした。

4.2 実装

大規模言語モデルとして、日本語 Llama3.1 モデルの Swallow-8B に指示チューニングを行ったモデル²⁾を使用した。実装には、vLLM³⁾ [18] を利用した。SamplingParams の temperature を 0.0, max_tokens を 512, stop を <|eot_id|> とし、その他の設定はデフォルト値とした。

言い換え応答の生成可能性を確認するために、LLM_{direct} と LLM_{2step} のそれぞれで応答を生成した。各手法では、学習データに含まれる言い換え応答と語りの組 667 個から、ランダムサンプリングを行い、Few-shot での応答生成時のプロンプトに含める事例を決定した⁴⁾。本実験では、shot 数を 0 から 10 まで変化させて、それぞれの生成性能を確認した⁵⁾。

プロンプトは、role と content をキーとする辞書型のデータのリストとして作成した。これを、Hugging Face の apply_chat_template メソッドを用いてテキストに変換して、大規模言語モデルに入力した。表2に、LLM_{direct} と LLM_{2step} に共通するシステムプロンプトを、表3に、LLM_{direct} が言い換え応答を生成するためのプロンプトをそれぞれ示す。また、LLM_{2step} が言い換え対象を選定するためのプロンプトと、言い換え応答を生成するためのプロンプトを表4と表5にそれぞれ示す。

2) <https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2>

3) <https://github.com/vllm-project/vllm>

4) ランダムサンプリングはシード値を固定して実施した。

5) LLM_{2step} の言い換え対象語句の選定における shot 数は 8 で固定し、著者が人手で正解を作成した事例を用いた。

表6 応答の生成例

語り	人間	LLM _{direct}	LLM _{2step}
城の名前は亀山です。えーかめは亀岡という名前は全国にえー数十か所ございますもんでえー	たくさん	亀山城は有名ですね。	多いんですね
んまー今の幸せな生活がああれあればそれでいいんじゃないかというふうに思います。	充分	それでいいんじゃないかと思いますね。	それでいいんじゃないかと思います。

4.3 評価方法

生成された応答を、人間が表出した応答との類似度によって評価した。評価指標には、BERTScore⁶⁾[19]を用いた。BERTScoreは、BERT[20]による埋め込み表現に基づいて類似度を計算する指標であり、BERTScoreが高い応答は、適切な言い換え応答であるとみなせる。BERTScoreは、適合率、再現率、F値の3種類存在するが、本実験ではF値を採用した。評価用データの各事例ごとに、BERTScoreのF値を計算し、その平均値を最終的な評価値とした。

4.4 実験結果

図3に各手法のBERTScoreを示す。ゼロショットのスコアは、LLM_{direct}が0.612、LLM_{2step}が0.665であった。最も高いスコアは、LLM_{direct}が0.704、LLM_{2step}が0.715であった。このことから、大規模言語モデルを用いることで、言い換え応答を一定の水準で生成できることを確認した。

LLM_{direct}とLLM_{2step}を比較すると、全てのshot数において、LLM_{2step}はLLM_{direct}よりもBERTScoreが高く、その中でもゼロショット時の性能差が最も大きかった。また、全てのshot数において、2つの手法の間に有意差が確認された($p < 0.05$)⁷⁾。これらのことから、言い換え応答の生成の前に言い換え対象を選定することの有効性が示唆された。

4.5 考察

言い換え応答の生成における誤りを分析した。LLM_{direct}とLLM_{2step}の10-shotでの設定で生成された応答に対して、言い換え応答とみなせる応答であるか否かを著者らが判断した。その結果、みなせないと判断された応答の割合は、LLM_{direct}では20.54%(53/258)、LLM_{2step}では9.69%(25/258)であった。LLM_{2step}では、言い換え応答でない応答を生成する誤りが減少していた。表6に応答の生成例を示す。1つ目の例では、LLM_{2step}は語りに含まれる

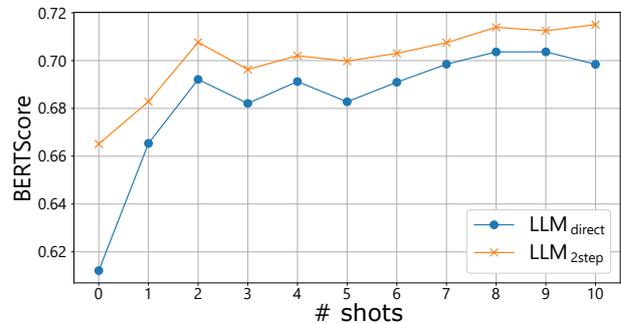


図3 LLM_{direct}とLLM_{2step}のBERTScore

「数十か所ございます」を「多いんですね」に言い換えることに成功している。一方で、LLM_{direct}が生成した応答は、語りに対する応答としては不自然ではないものの、言い換え応答とはいえない。

LLM_{2step}では、言い換え対象の選定と、選定した言い換え対象に基づく応答表現の生成にタスクを分解して、言い換え応答を生成している。これにより、言い換え応答表現の生成を担う大規模言語モデルが、表現の生成に注力することができ、言い換え応答とみなせない応答を生成する誤りが減少したものと考えられる。しかし、LLM_{2step}においても、そのような誤りは依然として含まれている。例えば、表6の2つ目の例のように繰り返し応答を生成する誤りがLLM_{2step}の出力に8個(25個)存在していた。

5 おわりに

本論文では、語りの傾聴のための言い換え応答の自動生成の実現に向けて、その表現の生成可能性を検証した。言い換え応答の生成実験を行い、大規模言語モデルに基づく生成手法により、一定の水準で言い換え応答の生成が可能であることを確認した。また、言い換え対象を選定したのち、その選定結果に基づいて言い換え応答を生成することで、生成性能が向上することを確認した。

本実験では、言い換え応答タイミングを既知として、応答表現の生成可能性を検証したが、今後は、言い換え応答タイミングの検出についても検討したい。

6) https://github.com/Tiiiger/bert_score

7) Wilcoxonの符号順位検定

謝辞

高齢者のナラティブコーパスは、奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室から提供いただいた。

参考文献

- [1] Labour Ministry of Health and Welfare. Summary report of comprehensive survey of living conditions 2022, 2023. https://www.mhlw.go.jp/english/database/db-hss/dl/report_gaikyo_2022.pdf.
- [2] Nigel Ward and Wataru Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics*, Vol. 32, No. 8, pp. 1177–1207, 2000.
- [3] Nicola Cathcart, Jean Carletta, and Ewan Klein. A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics*, Vol. 1, pp. 51–58, 2003.
- [4] 大野誠寛, 神谷優貴, 松原茂樹. 対話コーパスを用いた相づち生成タイミングの検出. 電子情報通信学会論文誌, Vol. J100-A, No. 1, pp. 53–65, 2017.
- [5] Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. BPM_MT: Enhanced backchannel prediction model using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3447–3452, 2021.
- [6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5998–6008, 2017.
- [7] Toshiaki Kawamoto, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. Generating repetitions with appropriate repeated words. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 852–859, 2022.
- [8] 石田真也, 井上昂治, 中村静, 高梨克也, 河原達也. 傾聴対話システムのための発話を促す聞き手応答の生成. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 77, pp. 1–6, 2016.
- [9] 井上昂治, ラーラーディベッシュ, 山本賢太, 中村静, 高梨克也, 河原達也. アンドロイド ERICA の傾聴対話システム-人間による傾聴との比較評価-. 人工知能学会論文誌, Vol. 36, No. 5, pp. H-L51_1–12, 2021.
- [10] 伊藤滉一朗, 村田匡輝, 大野誠寛, 松原茂樹. 語りの傾聴において不同意を示す応答の生成. 自然言語処理, Vol. 31, No. 1, pp. 212–249, 2024.
- [11] Kathleen McKeown. Paraphrasing questions using given and new information. *American Journal of Computational Linguistics*, Vol. 9, No. 1, pp. 1–10, 1983.
- [12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.
- [13] Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual LSTM networks. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2923–2934, 2016.
- [14] Sam Witteveen and Martin Andrews. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pp. 215–220, 2019.
- [15] Koichiro Ito, Masaki Murata, Tomohiro Ohno, and Shigeaki Matsubara. Construction of responsive utterance corpus for attentive listening response production. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pp. 7244–7252, 2022.
- [16] Eiji Aramaki. Japanese elder’s language index corpus v2, 2016. https://figshare.com/articles/dataset/Japanese_Elder_s_Language_Index_Corpus_v2/2082706/1.
- [17] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 日本語節境界検出プログラム CBAP の開発と評価. 自然言語処理, Vol. 11, No. 3, pp. 39–68, 2004.
- [18] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.
- [19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.
- [20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171–4186, 2019.