

音声対話における応答速度改善に向けた先読み技術の検討

大萩 雅也 水本 智也 吉川 克正

SB Intuitions 株式会社

{masaya.ohagi, tomoya.mizumoto, katsumasa.yoshikawa}@sbintuitions.co.jp

概要

本研究では大規模言語モデルを用いた音声対話システムにおける応答速度を改善するための新たな手法を提案する。既存のシステムではユーザー発話の終了後に音声チャットボットの応答を生成するため生成時間の分、応答開始が遅れていた。これに対し我々のシステムではユーザー発話が終了する前にそれまでの対話履歴から次にどのようなユーザー発話があるかを先読み予測し、それに対する応答を事前生成する。これにより生成時間を省略し予測ユーザー発話と実際のユーザー発話のマッチング時間のみで応答を返すことができるようになる。評価の結果、ユーザーの発話候補が限定される文脈のもとで我々の手法は高い先読み精度を発揮し応答速度の改善に寄与するという結果が得られた。

1 はじめに

大規模言語モデルの発達 [1]、また高性能な Speech2Text[2]、Text2Speech モデル [3] の開発により、自然な音声対話を行うことができるシステムが可能になりつつある。音声対話システムは幅広い応用先を持ち、今後社会に実装されることが期待される。しかしながら、大規模言語モデルを用いた音声対話システムの欠点の一つとして、応答生成に時間がかかるという問題が存在する。一般的に日本語話者は1秒以内に相手の発話に対して応答を始めるという傾向があるが [4]、例えば OpenAI の GPT-3.5-turbo は1秒では14トークンしか生成できず [5]、1秒以内に全ての生成を終えて応答を行うのは簡単でない。

既存の解決策として、言語モデルが全ての文章を生成し終える前に生成されたトークンを次々と音声として出力していくことが考えられる。しかしながら全て生成し終わる前に音声として発話を行う手法は有害な文章を弾くようなフィルタリングの適用が難しく、有害性が発見された時にはすでに文章の途中まで発話してしまっているような状況が起こ

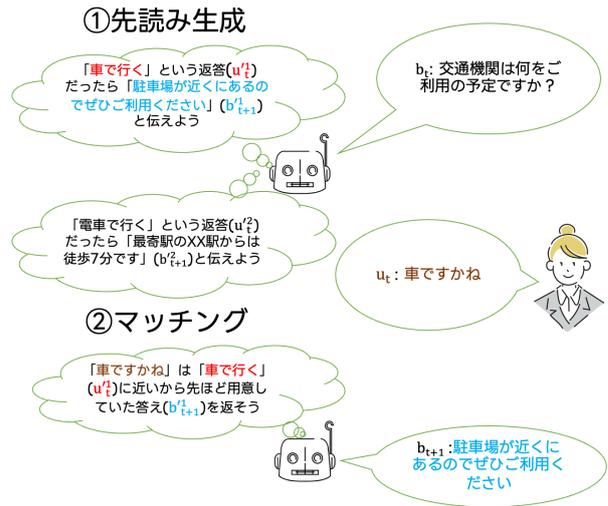


図 1: 音声発話をしながら同時にユーザーの発話を予測し、それに対するボットの応答を先に生成しておくことにより応答までの時間を短縮する

りうる。よって今回我々は発話を生成し終えてから音声出力を行う音声対話システムを対象とし、そこに発話の先読み技術を組み込むことで応答速度を改善することを提案する。

提案手法の概略図を図 1 に示す。既存の音声対話システムでは、ターン t においてボットの発話 b_t とユーザーの発話 u_t が行われている間、システムは b_t の音声出力を行いそして u_t が終了するのを待つだけである。我々の手法ではこの Idle な時間をユーザーの発話予測とボットの応答の事前生成に当てることで応答速度の改善を目指す。我々の手法ではそれまでの対話履歴からユーザー発話を複数パターン先読み予測し、パターンごとにボットの応答を事前生成する。その後、実際のユーザー発話 u_t が終わった後に予測ユーザー発話とのマッチングを行い、一定の類似度を越えた予測発話に関しては、先読みが成功したとして事前生成しておいたボットの発話を返す。この手法により、既存のシステムでは必然的にかかっていたボットの応答生成時間を予測発話とのマッチング時間のみで短縮することができる。

我々の先読みシステムを評価した結果、ユーザーの発話のおよそ 20%を高い類似度で先読みでき、マッチング時間のみで自然な応答を返すことができることが確認された。さらに詳細な分析を行った結果、直前のボットの発話がユーザーの発話を絞り込めるような文脈である時に先読み予測が成功しやすいということが判明した。

2 提案手法

我々のシステムは、ユーザーの次の発話を予測しそれに対する応答を準備しておく先読みモジュールと、予測ユーザー発話がどれくらい実際のユーザー発話と近かったかを検証し近い場合は準備した応答を返すマッチングモジュールの二つから構成される。各モジュールについて詳細に説明する。

まず先読みモジュールは対話のターン t においてそれまでの対話履歴から次のユーザー発話 u_t がどのような発話であるかを N パターン予測する。各予測ユーザー発話を u_t^k とする ($k=1,2,\dots,N$)。そして各 u_t^k に対するボットの応答 b_{t+1}^k を事前生成することで各予測に対してボットの応答を予め準備しておく。この予測と事前生成はボットが発話 b_t を音声出力し、ユーザーがそれに対する発話 u_t を終えるまでに完了することが期待される。そのため、ユーザーが u_t を発話し終えた時点で u_t^k が正しくユーザー発話 u_t を予測できているならば、それに対するボットの発話 b_{t+1}^k はすぐに応答として返すことができる状態であると言える。本研究では GPT-3.5-turbo を用いて few-shot を与えた上でユーザー発話予測とボットの応答事前生成を同時に行った。具体的な prompt は Appendix B に示す。各 u_t^k は次のマッチングモジュールで使うために sentence-bert を用いて埋め込み $e_{u_t^k}$ へと変換しておく。

次のマッチングモジュールでは実際のユーザー発話 u_t の終了後に、その埋め込み e_{u_t} と各予測ユーザー発話 $e_{u_t^k}$ とのコサイン類似度 $score_k$ を測る。この時、最大類似度を $score_{max}$ 、最大類似度を持つ先読み予測を u_t^{max} とする。この最大類似度が一定の閾値 T (今回は 0.75) を超えていた場合は先読みが成功したとみなし、事前生成しておいたボットの発話 b_{t+1}^{max} を返す。一定の閾値以下だった場合は通常通り生成を行う。マッチング時には u_t の埋め込みを作成し、予測ユーザー発話の埋め込みとのコサイン類似度を測るだけなので、言語モデルによる応答生成と比べて微小時間で終わらせることが可能であ

る。これにより、大規模言語モデルの文脈に応じた応答生成能力を活かしながら、素早い応答を実現することが可能となる。

3 実験

3.1 実験データ

実験用のデータとして、対話ロボットコンペティション 2022 [6] に出場した際に我々が収集した対話ログを用いた。この対話はボット側が旅行代理店の店員、ユーザー側がそこを訪れた客という設定のもと、ユーザーの旅行プランを決定するために 5 分間の対話を行うタスク指向型対話である。実際のボットと人間の会話を元にしたデータであり今回のシステムの実用性を測る上で効果的なデータであるため選定した。7 個の対話ログにおいてユーザー側の発話は合計 71 個存在し、それら全てに今回の先読みシステムを活用し、その評価を行う。

3.2 評価指標

先読みによる応答生成速度改善の評価を行うために、我々は 3 種類の評価指標を用いる。まず一つ目の指標は先読みにかかる時間に関する指標である。具体的な評価指標としては、先読みを生成し始めてから N パターン生成し終えるまでにどれほどの時間がかかるかを調べる look ahead duration, ユーザーの発話と先読みのマッチングにどれほどの時間がかかるかを調べる matching duration の二つを調べる。

なお、参考指標として今回の実験データを人手でアノテーションした結果、ボットの平均発話時間は 20.77 秒、ユーザーの発話も含めた平均発話時間は 23.90 秒であった。また、タスクの性質上観光地に関する説明を 30 秒以上時間をかけて行う場合があり、これらを省いて比較的短時間の発話のみに絞った場合のボットの平均発話時間は 11.21 秒、ユーザーの発話も含めた平均時間は 14.05 秒となった。

二つ目の指標は先読みの精度に関する指標である。この指標は二つに分けられる。一つ目はそれまでの対話履歴からどの程度 u_t を正確に予測できるかである。この指標は閾値を超えたデータの数 (一致数) と各データに対する平均最大類似度で計測する。二つ目は u_t^k と u_t の一致判定をどの程度正確に行えているかの一致度である。不正確な先読み発話を実際のユーザー発話と一致しているとみなしてしまうと、間違った文脈のもとでのボット発話が応答

として返ってしまう。この発生頻度を評価するために、最大類似度が閾値を超えていた先読みが、実際のユーザー発話とどの程度一致していたかを1(全く一致しない)から5(完全に一致している)の5段階で人手評価し、その平均を「一致度」とする。

三つ目の指標は事前生成したボットの応答がどの程度それまでの文脈に沿った応答になっているかである。 u_t^{max} が u_t を正確に先読みできていたとしても、それに次ぐ b_{t+1}^{max} が文脈から外れていれば先読みのメリットは失われてしまう。そのため、それまでの文脈から考えて b_{t+1}^{max} が自然なものになっているかどうかを、(0: 不自然, 1: 自然) の2段階で人手評価し、その平均を自然さとする。一致度、自然さに関しては主観評価であるため、アノテーター三人の評価の平均をとった値を報告する。

4 結果

表1に実験結果を示す。まず先読みにかかる時間に関する指標を見る。1パターンごとにおおよそ1秒ほどの生成時間が必要であり、予測パターン数が増えれば増えるほど先読みにかかる時間は増える。一方でマッチングにかかる時間は、パターン数が増えても所要時間はほぼ変わらない。また先読みにかかる時間に比べて無視できるほどの時間しかかかっていない。なお、Section 3.2に記載したように、ユーザーの発話も含めた平均発話時間は14.05秒であり、平均的にはN=20まで今回の手法は適用可能である。しかしながら図2にあるように、実際は5秒から10秒の発話が多く分布しているため、実用上はN=3,5が有用と考えられる。

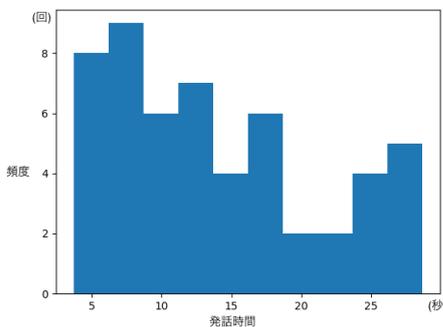


図2: ユーザー発話も含めた発話時間の分布

次に先読み精度を見る。まず一致数をみると、全ての設定において全体のユーザー発話71個中の16個、およそ20%以上を閾値以上の類似度で予測できている。我々の手法の一定の有効性を示している。

次に平均最大類似度を見ると、パターン数が増えるごとに類似度も上がっていていることが見て取れる。これはパターン数が増えるとそれだけ多くの回答の可能性をカバーできることを示唆しているが、平均最大類似度と一致数の間に強い相関は見られず、高い類似度での先読みはパターン数を増やしたからといって必ずしも増えるわけではないことがわかる。一致度については、おおよそ4に近いスコアを出しており、一致判定が行われたものに関しては基本的には正しく一致判定が取得している。

最後に、応答の自然さについて見ると、N=3においてはかなり高い割合で自然な応答を返している。不自然な応答を詳細に見てみると、一致判定の間違いがその後の不自然な応答にもつながっていることが見てとれた。例えば「質問があればお答え出来ませんが、いかがでしょうか?」というボットの発話に対して実際のユーザー発話が「トリックアート迷宮館ってどこにあるんですか」なのに対して、「トリックアート迷宮館の方が楽しそうだな。」という先読み発話を一致していると判断した結果、その後のボット発話が不自然なものになったデータが挙げられる。一致度が比較的高いN=3, N=10, N=15の自然さが高く、一致度が低いN=5, 20の自然さが低いことから一致度が自然さと相関するとわかる。

5 Case Analysis

先読み成功したものと失敗したものとの違いを詳細に分析するため、71個の全データの中で最大類似度が最も高かった1つと最も低かった2つを表2に示す。

まず最大類似度が最も高かったものを見てみると、「誰と観光するか」に関する問いがボットの最後の発話であり、ユーザーの応答の候補がある程度絞り込める文脈であるといえる。そこに回答としての典型例と呼べるような応答が返ってきており、これが先読み精度の高さにつながったことがわかる。次に最大類似度が最も低かった例をみると、まず一つ目の例はユーザーに対してQ&Aを投げかけるユーザーの応答が絞り込みにくい文脈であり、予測をするのが難しくなっている。さらに、そこからのユーザーの応答は直前のボットの発話のハルシネーションを指摘するものであり、こちらもボット側で予測するのは難易度が高い。次に二つ目の例を見てみると、こちらは文脈自体は「車の利用があるかどうかを聞く」という応答を絞り込みやすいものに

表 1: 先読みの評価結果

パターン数	時間		先読み精度			自然さ
	look ahead duration	matching duration	一致数	平均最大類似度	一致度	
N=3	3.99	0.0079	16/71	0.59	4.2/5	0.90
N=5	6.29	0.0080	18/71	0.62	3.9/5	0.72
N=10	9.28	0.0081	16/71	0.63	4.1/5	0.79
N=15	12.11	0.0091	20/71	0.63	4.1/5	0.78
N=20	13.6	0.0092	17/71	0.64	3.8/5	0.76

表 2: 最大類似度が最も高かったもの 2 つと最も低かったもの 2 つ

対話履歴	正解ユーザー発話	最大類似度
<p>ボット: ところで、本日は東京臨海広域防災公園と、東京都水の科学館のどちらに行くかで、迷っていらっしやるとお伺いしました。ちなみに、どちらか今までに行かれたことがありますか？</p> <p>ユーザー: 両方行ったことがないので選びました</p> <p>ボット: 簡単にご紹介いたしますと東京臨海広域防災公園は首都圏防災の拠点施設を見学できる施設です。東京都水の科学館は水の大切さを科学の視点で紹介する施設です。どちらのスポットが良いか考えたいので、簡単にお客様の情報をお聞かせくださいね。まずですが、今回はどなたと一緒に観光されますか？</p>	一人でいきます	0.999
<p>ユーザー: 今日のお天気はどうか</p> <p>ボット: 今日のお天気ですね。検索によると本日の東京地方の天気は晴れ時々曇りでしょう。最高気温は 24° C、最低気温は 13° C となりそうです。他に質問はございますか？</p>	24° は嘘じゃないですか	0.277
<p>ボット: 遊園地などテーマパークはお好きですか？</p> <p>ユーザー: そんなに行かないけど友達が行くって言ったら渋々行くかな</p> <p>ボット: 私も実はあまり得意ではないのですが、せっかくなので行ってみようと思います。それから交通手段なんですけども、お車のご利用はありますか？</p>	ペットの月島くんが暑いのが苦手です涼しい環境がいいです	0.246

なっているが、ユーザー側の発話が「車の利用があるかどうか」に対する直接の回答ではなく、「ペットに配慮した交通手段にしてほしい」という婉曲的な回答となっているため、予測するのが難しくなっている。これらの結果から、先読み精度はユーザー発話候補が絞り込みやすい文脈のもとで高くなりやすいが、その候補から外れた回答が返ってくると低くなってしまふことがわかる。

6 関連研究

対話システムは大きく分けてルールベースのもの [7] と生成モデルベースのもの [8] が存在する。生成モデルベースのものはモデルが大規模である程計算時間がかかるため応答にかかる時間が問題となる [9, 10]。我々はその解決策として先読みを提案した。

対話の先読みに関する既存研究は複数存在する。例えば先読みを行うことで、タスクの効率の良い解決を目指すもの [11, 12, 13] や、他タスクの解決のために先読みを補助情報として活用するもの [14, 15] などが上げられる。しかしながら、本研究のようにリアルタイム音声対話における応答生成時間の短縮を目的として先読み技術を導入した研究は我々の知る限り存在しない。本研究ではリアルタイム性に対応するために、ユーザー発話が終了するまでに先読

みを終わらせるという制約や実際のユーザー発話と予測ユーザー発話とのマッチングといった新たな課題に取り組んでいる。この点において、本研究で取り組んでいるタスクは既存の対話先読み技術では取り組まれてこなかった新たなものであると言える。

7 結論

本研究では、対話履歴から次のユーザーの発話とそれに対するボットの応答を先読み生成することで大規模言語モデルを活用した音声対話システムにおける応答速度の改善手法を提案した。我々が提案した評価指標のもとシステムを評価した結果、ユーザーの発話のおよそ 20% を高い類似度で先読みすることができ、応答速度改善において一定の効果を発揮することがわかった。しかしながら、先読み結果の誤った一致判定により不自然な発話を返してしまうこともあったため、一致判定の性能を高めていくことが重要である。将来的な方向性としては、実際のユーザーに先読みを搭載した対話システムを試してもらい、先読みが搭載されていないシステムとの比較を行うことが挙げられる。さらに、先読みコストを下げるために対話履歴から事前に先読みをするべきかどうかを判別することなども挙げられる。

参考文献

- [1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, **Advances in Neural Information Processing Systems**, 2022.
- [2] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In **Proceedings of the 40th International Conference on Machine Learning**, ICML'23. JMLR.org, 2023.
- [3] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai, 2023.
- [4] Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. Universals and cultural variation in turn-taking in conversation. **Proceedings of the National Academy of Sciences**, Vol. 106, No. 26, pp. 10587–10592, 2009.
- [5] OpenAI Developer Forum. Gpt-3.5 and gpt-4 api response time measurements, 2023. <https://community.openai.com/t/gpt-3-5-and-gpt-4-api-response-time-measurements-fyi/237394>.
- [6] Ryuichiro Higashinaka, Takashi Minato, Hiromitsu Nishizaki, and Takayuki Nagai. Proceedings of the dialogue robot competition 2022, 2022.
- [7] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. **Commun. ACM**, Vol. 9, No. 1, p. 36–45, jan 1966.
- [8] Iulian V. Serban, Alessandro Sordani, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In **Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence**, AAAI'16, p. 3776–3783. AAAI Press, 2016.
- [9] Takato Yamazaki, Tomoya Mizumoto, Katsumasa Yoshikawa, Masaya Ohagi, Toshiki Kawamoto, and Toshi-nori Sato. An open-domain avatar chatbot by exploiting a large language model. In **Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 428–432, 2023.
- [10] 川本稔己, 山崎天, 坂田亘, 佐藤敏紀. Hyperclova を利用したプロンプトプログラミングによるシチュエーションに適した応答生成. 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. 93, pp. 131–136, 2021.
- [11] Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 3014–3026, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [12] Yosuke Kishinami, Reina Akama, Shiki Sato, Ryoko Tokuhisa, Jun Suzuki, and Kentaro Inui. Target-guided open-domain conversation planning. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 660–668, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
- [13] Zhuoxuan Jiang, Xian-Ling Mao, Ziming Huang, Jie Ma, and Shaochun Li. Towards end-to-end learning for efficient dialogue agent by modeling looking-ahead ability. In Satoshi Nakamura, Milica Gasic, Ingrid Zuckerman, Gabriel Skantze, Mikio Nakano, Alexandros Papangelis, Stefan Ultes, and Koichiro Yoshino, editors, **Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue**, pp. 133–142, Stockholm, Sweden, September 2019. Association for Computational Linguistics.
- [14] Eyal Ben-David, Boaz Carmeli, and Ateret Anaby-Tavor. Improved goal oriented dialogue via utterance generation and look ahead, 2021.
- [15] Shuo Zhang, Junzhou Zhao, Pinghui Wang, Yu Li, Yi Huang, and Junlan Feng. “think before you speak” : Improving multi-action dialog policy by planning single-action dialogs. In Lud De Raedt, editor, **Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22**, pp. 4510–4516. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

A 最大類似度の分布

図3にN=20における最大類似度の分布図をプロットした。大まかに0.6から0.7を中心とした正規分布となっていることが見て取れる。ここからは、予測しやすい発話と予測しづらい発話は均等に分布しており、その中間に属する発話に関する先読み精度をどれほど上げていけるかが今後の先読み精度の向上に必要な不可欠であるということがわかる。

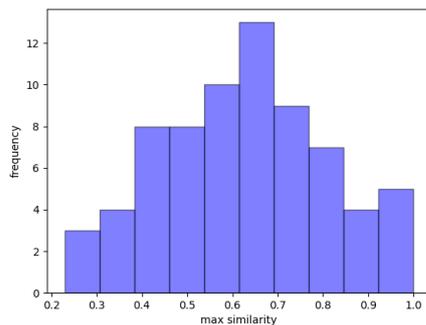


図3: 最大類似度分布

B 先読みモジュールのプロンプト

```
### 指示
次のお客様とショウコの「会話履歴」をもとに、次のお客様の発話とそれに対するショウコの発話を「制約条件」に従うように10パターン予測してください

### 制約条件
「出力」は「数字: お客様: xx, ショウコ: yy\n」の形で順番に出力してください。
以下の「例: 会話履歴」と「例: 出力」を参考にして出力してください

### 例: 会話履歴
ショウコ: 本日はよろしくお願いたします。歴史的な施設をご覧になることは好きですか?
### 例: 出力
1: お客様: 好きです。 , ショウコ: 承知しました。お寺や神社などの歴史的な施設を中心に紹介しますね。
2: お客様: いや、あまり好きではないですね。 , ショウコ: 承知しました。では自然が楽しめる場所や美術館などの施設をご紹介しますね。
3: お客様: どちらでもないですかね。 , ショウコ: なるほど。でしたら他の観光地を楽しんだ上でお時間があれば回るのが良さそうですね。

### 会話履歴
ショウコ: いらっしゃいませ。ショウコと申します。よろしくお願いたします。
お客様: よろしくお願いたします
ショウコ: 私は観光案内の仕事をしているのですが、お客様はどのようなお仕事をされていますか?
お客様: 編集者です
ショウコ: 具体的にはどのようなことをされているのですか?
### 出力
```

図4: 先読み生成のためのプロンプト