音声対話システムのための意味的類似度を考慮した予測信頼度モデル

森 清忠 [1,2] 河野 誠也 [2,1] アンケル ガルシア コントレラス [2] 吉野 幸一郎 [3,2,1]
[1] 奈良先端科学技術大学院大学
[2] 理化学研究所ガーディアンロボットプロジェクト [3] 東京科学大学
mori.kiyotada.mh5@naist.ac.jp
{seiya.kawano,angel.garciacontreras,koichiro.yoshino}@riken.jp

概要

一般的な音声対話システムはユーザの発話終了検知後に音声応答を生成するため，ユーザが応答生成の時間に待機するユーザ知覚遅延（UPL）が発生する．予測信頼度モデルは，予測したユーザ発話と完全なユーザ発話の文字列が一致する確率を推定するもので，UPL を軽減するために考案された．しかしシステムがユーザ発話を完全に予測せずとも，発話の途中で発話全体の意味を捉えることで，適切な対話応答を生成できる可能性がある．本研究では予測信頼度モデルを再定義し，発話同士の意味的類似度を推定させることで，応答の精度を維持しながら UPL をさらに削減できることを示す．

1 はじめに

カスケード型音声対話システムは 3 つの手順で音声応答を生成する [1]．まず音声認識モデルがユーザの音声発話を書き起こす．次に対話モデルが書き起こしから応答文を生成する．最後に音声合成モデルが応答文から音声応答を生成する．この一連の流れにおいて，音声対話システムはユーザの発話終了検知後に音声応答を生成するため，発話終了検知と応答生成の時間に待機が生じる．これは人間同士の対話では生じない遅延であり，対話システムの自然さの低下に繋がる [2]．

この遅延を削減する手法として音声応答のプリフェッチング [3, 4] が提案されている．その中でも，ユーザの音声発話を段階的に音声認識しながら，完全なユーザ発話を先読みし，その先読みに対して事前に音声応答を生成する手法がある．プリフェッチングが成功すれば，システムは発話終了検知後にすぐさま応答できる．一方で，プリフェッチングしたシステム応答はユーザの部分発話に対して用意されたもので，完全なユーザ発話に対して用意したものと常に同等のパフォーマンスが実現できるわけではない．

こうした音声対話におけるプリフェッチングの問題は，ユーザ発話の中途で後続しうるユーザ発話に含まれる文字列を適切に予測可能かどうかという問題として定式化されている．しかし，人間同士の対話を考えてみると，人間は相手の発話を一言一句正確に捉えて対話応答を生成しているわけではない [5]．プリフェッチング可能かどうかはユーザ発話の意図を正しく把握し適切な対話応答を準備できるかに依存する部分が大きい．

そこで本研究では，発話が持つ意味の観点から予測信頼度モデルを再定義し用いることを試みる．具体的には，ユーザの部分発話の先読みと完全なユーザ発話のテキストの一致を測るのではなく，意味的類似度を測ることで予測信頼度モデルの学習を行う．これにより，従来の予測信頼度モデルよりも遅延を削減しながら，プリフェッチングした応答精度を人間が違和感を感じない程度に維持することを試みる．

再定義した予測信頼度モデルを，対話応答のプリフェッチングのタイミングと，プリフェッチングした対話応答の応答精度の観点から評価する．その結果，ユーザ発話同士の意味的類似度が閾値以上である確率を推定する予測信頼度モデルは，従来の予測信頼度モデルと比べ，早いタイミングでの対話応答のプリフェッチを可能とすることがわかった．また意味的類似度の適切な設定により，プリフェッチングしたシステム応答と完全なユーザ発話に対する応答をそん色なくできることが示唆された．実験は英語のタスク対話データセットである MultiWOZ [6]

と，その日本語版である JMultiWOZ [7] で行った．本研究では，音声認識精度が十分であると仮定し，実際の音声認識結果ではなくテキストで実験した．

我々の貢献は以下の通りである．

1. 予測したユーザ発話と完全なユーザ発話の意味的類似度が閾値以上である確率を推定する予測信頼度モデルの構築と評価
2. 英語と日本語で訓練したそれぞれの予測信頼度モデルの性能評価の言語差の分析

2 関連研究

2.1 Personalized Predictive ASR

Personalized Predictive ASR [8] は，音声発話をプリフェッチングし，ユーザ知覚遅延 (UPL: User Perceived Latency) を削減するパイプライン手法を提案した．UPL の定義を式 (1) に示す．

$$T_{\mathrm{UPL}} = \begin{cases} \max(T_{\mathrm{EP}}, T_{\mathrm{PF}} + T_{\mathrm{Response}}) & \text{(Successful)} \\ T_{\mathrm{EP}} + T_{\mathrm{Response}} & \text{(Failed)} \end{cases} \quad (1)$$

T_{EP} はシステムがユーザ発話終了後に発話終了を検知するまでの時間である．T_{Response} はシステムが音声応答を生成する時間である．T_{PF} は予測信頼度モデルが応答のプリフェッチングを決定してから，システムが音声応答を生成するまでの時間である．

対話応答のプリフェッチングの目標は，プリフェッチングした応答の精度を維持しながら，T_{UPL} を最小化することである．ここでユーザ発話終了の $\Delta T(\%)$ 手前で，任意のプリフェッチングの成功の定義が満たされる時，$T_{\mathrm{PF}} = -\Delta T$ と定義する．ΔT がこのプリフェッチによる時間利得，つまり予測利得である．すなわち，T_{UPL} の最小化は予測利得の最大化を意味する．

関連研究での発話予測モデルと予測信頼度モデルの定義をそれぞれを式 (2), (3) に示す．

$$\hat{y}_{\mathrm{full},t} \approx \underset{y_{\mathrm{full}}}{\operatorname{argmax}}\, P(y_{\mathrm{full}} | \hat{y}_t) \quad (2)$$

$$P(y_{\hat{\mathrm{full}},t} = y_{\hat{\mathrm{full}}}) \quad (3)$$

$\hat{y}_t$ は時刻 t でのユーザ発話である．$\hat{y}_{\mathrm{full},t}$ は $\hat{y}_t$ から $\hat{y}_{\mathrm{full}}$ の予測文である．y_{full} と $y_{\hat{\mathrm{full}}}$ とはそれぞれ完全なユーザ発話とそれに対応する音声認識の最終仮説文を示す．本研究では $y_{\hat{\mathrm{full}}} = y_{\mathrm{full}}$ である．

この定義を用いた場合の対話応答のプリフェッチング手順を以下に示す．ここで T. は条件が満たされた場合，F. は条件が満たされない場合とする．

1. ユーザの音声発話を段階的に音声認識する
2. 発話予測モデルで，音声認識した仮説文から完全なユーザ発話を予測する
3. 予測信頼度モデルで，予測したユーザ発話と音声認識の最終仮説文が一致する確率を推定する
4. 予測信頼度モデルが出力した確率が一定以上である場合，
   T. システム応答をプリフェッチングし，音声応答を生成する
   F. 手順 1 に戻る
5. プリフェッチング時に，予測したユーザ発話と音声認識の最終仮説文が一致する場合，
   T. プリフェッチングしたシステム応答でユーザ発話に返答する
   F. 完全なユーザ発話に対する，システムの音声応答を生成しユーザ発話に返答する

2.2 An Incremental Turn-Taking Model

An Incremental Turn-Taking Model [9] では，ユーザ発話の単語ごとのターンテイキングの可能性を推定する Incremental Turn-Taking Decider (iTTD) を提案した．iTTD の定義を式 (4) に示す．

$$P(\mathrm{take}_{\mathrm{turn}} | \mathrm{s}_t) = P(0 | \mathrm{s}_t) \quad (4)$$

ここで s_t は時刻 t での対話状態である．0 は s_t と完全なユーザ発話の対話状態が一致することを示す記号である．

iTTD は予測発話の一致ではなく部分発話から推定される対話状態の一致を用いている．このため，予測信頼度モデルの平均予測利得よりも高い傾向にある．それぞれの実験結果はテストデータに強く依存するため，正確な比較は困難であるが，予測信頼度モデルの平均予測利得は 0.23 であり，iTTD の平均予測利得は 0.61 であった．

この対話状態を用いたモデルの成功は，予測信頼度モデルのプリフェッチングの成功の定義を，発話の意味を考慮して適切に再定義することで，対話応答のプリフェッチングの予測利得を向上できることを示唆している．

3 提案手法

本研究では，予測信頼度モデルを先読みしたユーザ発話と完全なユーザ発話の意味的類似度が一定以上である確率を推定するものと再定義する．本研究

での予測信頼度モデルの定義を式 (5) に示す．

$$P\left(\text{S-BERT}(\hat{y}_{\text{full},t}, \hat{y}_{\text{full}}) > T\right) \tag{5}$$

ここで S-BERT は Sentence BERT (stsb-xlm-r-multilingual) [10] での意味的類似度の計算である．T は任意の閾値である．

この予測信頼度モデルは，BERT (bert-base-multilingual-uncased) [11] の CLS ベクトルに対するファインチューニングに以下の実験で定義するラベルを用いることで構築する．この再定義に対して実験で確認すべき点は以下の 2 つである．

1. 最適な閾値 T はどのような値になるのか
2. 新しい予測信頼度モデルは完全なユーザ発話への応答とそん色ない応答をプリフェッチング可能かどうか

4 実験

4.1 データセット

MutliWOZ と JMultiWOZ で，それぞれ英語と日本語の予測信頼度モデルの訓練と評価を行う．以下に訓練データセットの構築手順を示す．日英の異なりは，時刻 t の単位として文字を用いるか単語を用いるかである．

1. MultiWOZ のユーザの発話文をそれぞれ単語単位で区切る
2. 各単語ごとの段階的なユーザ発話に対して，大規模言語モデルを用いた発話予測モデルでユーザ発話の先読みを 1 つ生成する
3. 各単語ごとの段階的なユーザ発話に対して，発話予測モデルで，生成したユーザの先読み発話に対するシステム応答を 4 つ生成する

ここで英語と日本語の発話予測モデルは Qwen (Qwen2.5-14B-Instruct) [12] をそれぞれ MutliWOZ と JMultiWOZ の Validation データセットで Low-Rank Adaptation (LoRA) [13] により微調整したものを用いる．その際の入力は，対話状態，過去最大 4 つの対話履歴，部分発話，1 つの応答例とするアルパカプロンプトである．対話履歴がないものは予測が困難になるため利用していない．この時，ハイパーパラメータはエポック数が 1，学習率が $2e-4$，バッチサイズが 32，LoRA ランクが 16，最適化関数が Adam 8bit，最大入力トークン数が 2048，temparature が 1 である．

予測信頼度モデルの訓練データセットの要素を以下に示す．

- $\hat{r}_{\text{full},t}$: $\hat{y}_{\text{full}} \neq \hat{y}_t$ の際の，$\hat{y}_{\text{full},t}$ への発話予測モデルによるシステム応答
- $\hat{r}_{\text{full}}$: $\hat{y}_{\text{full}}$ への発話予測モデルによるシステム応答
- r_{full}: $\hat{y}_{\text{full}}$ に対するデータセットに含まれるシステム応答
- h_{dialogue}: $\hat{y}_{\text{full}}$ 以前の最大過去 4 発話

4.2 ラベル

予測信頼度モデルの訓練及び評価のため $l_{\text{sbert}T}$ と l_{literal} の 2 種類のラベルを生成した．ここで $l_{\text{sbert}T}$ は $\mathit{S\text{-}BERT}(\hat{y}_{\text{full},t}, \hat{y}_{\text{full}}) => T$ の時に正例，そうでない時に負例となるラベルである．本実験では $T = \{0.75, 0.80, 0.85, 0.90, 0.95\}$ である．一方，l_{literal} は $\hat{y}_{\text{full},t} = \hat{y}_{\text{full}}$ の時に正例，そうでない時に負例となるラベルである．このラベルは従来の予測信頼度モデルと同様の予測成功の定義に基づいている．

4.3 訓練

英語と日本語の予測信頼度モデルは，BERT をそれぞれ MutliWOZ と JMultiWOZ の Test データセットの 50 対話 (対話番号 0-49) で微調整したものを用いる．予測信頼度モデルの入力特徴量は h_{dialogue}，$\hat{y}_t$，$\hat{y}_{\text{full},t}$ である．予測信頼度モデルの訓練時のハイパーパラメータはエポック数が 1，学習率が $2e-5$，バッチ数が 16，損失関数は Focal Loss ($\gamma = 2.0$) [14]，出力層は Softmax 関数である．

4.4 評価

英語と日本語の予測信頼度モデルの評価は，それぞれ MutliWOZ と JMultiWOZ の Test データセットの 25 対話 (対話番号 50-74) のユーザ発話で評価した．今回構築した予測信頼度モデルを評価するためには，予測信頼度モデル学習の成否，予測信頼度モデルが得た予測利得，予測信頼度モデルによってプリフェッチングしたシステム応答群の適切性の 3 点を評価する必要がある．まず，予測信頼度モデル学習の成否については以下の 3 指標で評価する．

- Successful Prefetch Rate (SPR): 予測信頼度モデルが初めて予測成功と判断した時に，実際にプリフェッチングの成功の定義を満たす割合
- Failed Prefetch Rate (FPR): 予測信頼度モデルが初

表 1 予測信頼度モデルの予測・応答評価

Model	SPR ↑	FPR ↓	NPR	P-Gain ↑	C-Gain ↑	Total ↑	Comp	ROUGE ↑	S-BERT ↑	PR > R ↓
EN-PCM$_{l_{\text{sbert075}}}$	0.28	0.21	0.51	0.44	20.18	85	0.98	0.51	0.65	**0.48**
EN-PCM$_{l_{\text{sbert080}}}$	0.26	0.23	0.51	0.43	18.51	74	1.00	0.51	0.67	**0.45**
EN-PCM$_{l_{\text{sbert085}}}$	0.20	0.29	0.51	0.37	15.95	60	0.97	0.58	0.74	**0.46**
EN-PCM$_{l_{\text{sbert090}}}$	0.17	0.31	0.51	0.34	14.08	51	0.96	0.64	0.76	**0.40**
EN-PCM$_{l_{\text{sbert095}}}$	0.17	0.32	0.51	0.24	9.36	48	0.98	0.69	0.80	0.52
EN-PCM$_{l_{\text{literal}}}$	0.14	0.30	0.56	0.07	2.29	34	1.00	0.86	0.93	0.56
JA-PCM$_{l_{\text{sbert075}}}$	0.30	0.19	0.51	0.71	20.58	120	0.86	0.61	0.69	0.57
JA-PCM$_{l_{\text{sbert080}}}$	0.32	0.18	0.50	0.66	18.31	138	0.81	0.70	0.77	0.52
JA-PCM$_{l_{\text{sbert085}}}$	0.28	0.22	0.50	0.61	15.90	130	0.75	0.73	0.81	**0.45**
JA-PCM$_{l_{\text{sbert090}}}$	0.26	0.23	0.51	0.57	14.50	122	0.75	0.77	0.82	**0.40**
JA-PCM$_{l_{\text{sbert095}}}$	0.22	0.28	0.51	0.50	12.53	104	0.72	0.78	0.82	**0.41**
JA-PCM$_{l_{\text{literal}}}$	0.19	0.31	0.51	0.23	5.03	88	0.72	0.89	0.94	0.66

めて予測成功と判断した時に，プリフェッチングの成功の定義を満たさない割合

- Non Prefetch Rate (NPR): 予測信頼度モデルが発話の終了まで予測成功と判断しない割合

次に，予測信頼度モデルが得た予測利得については以下の 2 指標で評価する．

- Prediction Gain (P-Gain): プリフェッチング成功時の平均予測利得
- Character Gain (C-Gain): プリフェッチング成功時の $\hat{y}_t$ と $\hat{y}_{\text{full}}$ の文字数の差

最後に，プリフェッチングしたシステム応答群の評価は以下の 5 指標で行う．ここで，応答順位モデルとして Athena-RR [15] を用いた．

- Total: プリフェッチングの成功回数
- Comp: プリフェッチング成功時の $r_{\text{full}} \neq \hat{r}_{\text{full},t}$ の割合
- S-BERT: $r_{\text{full}} \neq \hat{r}_{\text{full},t}$ の際の，$\hat{r}_{\text{full},t}$ と 4 つの $\hat{r}_{\text{full}}$ の最大の意味的類似度
- ROUGE [16]: $r_{\text{full}} \neq \hat{r}_{\text{full},t}$ の際の，$\hat{r}_{\text{full},t}$ と 4 つの $\hat{r}_{\text{full}}$ の最大の ROUGE-1 の F1 スコア
- PR > R: Athena-RR が h_{dialogue}，$\hat{y}_{\text{full}}$，r_{full} を入力として，$\hat{r}_{\text{full},t}$ よりも $\hat{r}_{\text{full}}$ を評価する割合

Athena-RR は英語のみに対応しているため，日本語の予測信頼度モデルの評価において，Argos Translate [17] で英訳したテキストを入力としている．

5 実験結果

評価結果を表 1 に示す．ここで EN/JA はデータセットの言語，sbert*の文字列は予測信頼度モデルの学習に用いたラベルを表す．

表 1 より，$l_{\text{sbert}T}$ で微調整した予測信頼度モデルの P-Gain は，l_{literal} で微調整した予測信頼度モデルの P-Gain の約 2 倍である．また応答順位モデルは，適切な意味的類似度の閾値 T で訓練した予測信頼度モデルがプリフェッチングした応答群と実際の応答群を，区別できないことが示唆された．実際，英語の予測信頼度モデルでは PR>R は，$T = \{0.75, 0.80, 0.85, 0.90\}$ の時，日本語の予測信頼度モデルでは，$T = \{0.85, 0.90, 0.95\}$ の時，0.50 以下である．

また予測信頼度モデルは，PR>R が 0.50 以下である T の範囲を見ると，日本語の方が英語よりも，プリフェッチングした応答群の精度を保つためにより高い意味的類似度の閾値を必要とする傾向があることが示唆した．

6 おわりに

本研究では，予測信頼度モデルで予測したユーザ発話と完全なユーザ発話の意味的類似度が一定以上である確率を推定した．その結果，プリフェッチングしたシステム応答の精度を維持しながら，ユーザ知覚遅延を大きく削減できることを示した．一方で，意味敵類似度等による応答の評価においてはプリフェッチングした応答が完全なユーザ発話に対する応答よりも劣化していることが示されている．

今後は，人間がプリフェッチングした応答を許容可能か人手評価実験を実施する．また提案手法を実際の音声対話システムに適応し，その実用性を評価する。

謝辞

本研究の一部は JST ムーンショット型研究開発事業 JPMJMS2236 の支援を受けたものです.

参考文献

[1] Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. Wavchat: A survey of spoken dialogue models. **arXiv preprint arXiv:2411.13577**, 2024.

[2] Derek Jacoby, Tianyi Zhang, Aanchan Mohan, and Yvonne Coady. Human latency conversational turns for spoken avatar systems. **arXiv preprint arXiv:2404.16053**, 2024.

[3] Hikaru Kamioka, Satoshi Maeda, and Masayuki Hashimoto. Response delay reduction in large language model-based dialogue systems. **Journal of Machine Intelligence and Data Science (JMIDS)**, Vol. 5, , 2024.

[4] Oswald Zink, Yosuke Higuchi, Carlos Mullov, Alexander Waibel, and Tetsunori Kobayashi. Predictive speech recognition and end-of-utterance detection towards spoken dialog systems. **CoRR**, 2024.

[5] Gabriel Skantze. Exploring human error recovery strategies: Implications for spoken dialogue systems. **Speech Communication**, Vol. 45, No. 3, pp. 325–341, 2005.

[6] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In **Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020**, pp. 109–117, 2020.

[7] Atsumoto Ohashi, Ryu Hirai, Shinya Iizuka, and Ryuichiro Higashinaka. Jmultiwoz: A large-scale japanese multi-domain task-oriented dialogue dataset. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9554–9567, 2024.

[8] Andreas Schwarz, Di He, Maarten Van Segbroeck, Mohammed Hethnawi, and Ariya Rastrow. Personalized predictive asr for latency reduction in voice assistants. In **INTERSPEECH 2023**, pp. 745–749, 2023.

[9] Andrei C. Coman, Koichiro Yoshino, Yukitoshi Murase, Satoshi Nakamura, and Giuseppe Riccardi. An incremental turn-taking model for task-oriented dialog systems. In **INTERSPEECH 2019**, pp. 4155–4159, 2019.

[10] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 11 2019.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. **CoRR**, Vol. abs/1810.04805, , 2018.

[12] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.

[14] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In **proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 2980–2988, 2017.

[15] Vrindavan Harrison, Rishi Rajasekaran, and Marilyn Walker. A transformer-based response evaluator for open-domain spoken conversation. **arXiv preprint arXiv:2302.04424**, 2023.

[16] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, 2004.

[17] Argos Open Tech. Argos translate, 2024. Accessed on Jan 9, 2025.