Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain - August 19 - 21, 2024
Paper No. MHCI 104
DOI: 10.11159/mhci24.104

Reducing Response Delays in Dialogue Systems Using the Predictive Performance of Large Language Models

Masayuki Hashimoto[1]
[1]Toyo University
2100 Kujirai, Kawagoe-shi, Saitama, Japan
hashimoto065@toyo.jp

Extended Abstract

1. Objectives

Current speech dialogue systems struggle to provide smooth, human-like responses to user utterances, often resulting in response delays. In this study, we propose an approach to reduce these delays by utilizing the predictive capabilities of Large Language Models (LLMs)[1] within speech dialogue systems. We conducted an analysis to estimate the potential reduction in response times that can be expected from this proposed approach. Furthermore, we propose methodologies to implement this approach in actual speech dialogue systems.

2. Approach

In typical LLM-based speech dialogue systems, the system waits for the user to complete their utterance before feeding it into the LLM to generate a response and process the speech. This causes a delay equal to the time taken by the LLM to generate the response text after the user has finished speaking. To reduce this delay, we propose an approach where the LLM generates response text for an incomplete utterance without waiting for the user to finish speaking. This allows for earlier initiation of LLM response generation, potentially reducing response delays. Additionally, the system verifies that the user has finished speaking before responding.

3. Scope

This study focuses on text sentences gained after speech recognition in speech conversation systems. While there are some deviations from actual spoken dialogues, a text chat corpus[2] was employed for its clear content when the proposed approach was evaluated in this study. Additionally, a corpus derived from transcriptions of real conversations[3] was also examined.

4. Issues

However, two significant issues arise with the proposed approach described in Section 2. First, despite LLMs' ability to predict the continuation of unfinished sentences, they may not always produce appropriate responses when generating replies to incomplete utterances. As part of a feasibility study, this research estimated the expected reduction in response time using gpt-3.5-turbo, as shown in Section 5. Second, in practical dialogue systems, it is unclear at which point during a user's ongoing utterance it would be appropriate to stop voice data acquisition and initiate dialogue response generation. This study proposes a metric to determine the timing for response generation that minimizes response failures, as shown in Section 6, and will quantitatively evaluate how much this proposed algorithm can reduce failure rates.

5. Anticipated Effects and Analysis Results / Expected Reduction in Response Time

To determine the predictive performance, we assessed gpt-3.5-turbo's ability to generate coherent responses from incomplete user utterances, even when part of the utterance was missing. Results showed that failure rates increase with the number of omitted characters. Specifically, in an evaluation with a text chat corpus where 15 characters were omitted, denoted as $N= 15$, the failure rate was under 20%. Translating 15 characters to about 2.5 seconds of spoken Japanese suggests that it may be possible to reduce response times by up to 2.5 seconds with an 80% success rate. In contrast, a spoken dialogue corpus showed a 47% failure rate at $N = 15$. For this issue, a reevaluation is planned using a more exemplary daily conversation corpus.

6. Proposed Method Based on Experimental Results

As a criterion for deciding when to stop voice data acquisition and initiate dialogue response generation, we focus on the embedding representations of dialogue content using Sentence-BERT[4]. This approach is inspired by the fact that LLMs typically use the embedding representation of an input sentence to generate subsequent words. We investigated changes in the similarity measure, $S_t(X_0,X_t)$, between the embedding representation X_0 (the embedding at the start of the user's utterance) and X_t (the embedding at midpoint t during the user's utterance). Our analysis aimed to determine whether LLMs can produce coherent responses based solely on the dialogue history up to t. We discovered that when S_t sharply declines and then rises, gpt-3.5-turbo is capable of returning a coherent response based solely on the dialogue history up to that point. Consequently, our proposed method involves continuously transcribing the user's speech in real-time, calculating S_t from the entire dialogue history up to that point, and initiating dialogue response generation when the characteristic temporal change in S_t —a sharp decline followed by a rise—is detected. Further analysis and quantitative evaluation of this timing detection method are planned.

References

[1] Long Ouyang, JeffreyWu, 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," Advances in Neural Information Processing Systems, 2022.

[2] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro, "Empirical Analysis of Training Strategies of Transformer-based Japanese Chitchat Systems." arXiv:2109.05217. (2021)

[3] Itsuko Fujimura, Shoju Chiba, Mieko Ohso, "Lexical and Grammatical Features of Spoken and Written Japanese in Contrast:Exploring a lexical profiling approach to comparing spoken and Written corpora," in *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*, 2012, 393-398.

[4] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982-3992.