5ZM-07

Improving the communication methods of GPT based on the language proficiency of learners

Zihan Zhang† Shin'ichi Konomi‡
†Graduate School of ISEE, Kyushu University
‡Faculty of Arts and Science, Kyushu University

1. Introduction

With the rise and widespread use of large-scale language models such as GPT, interactive foreign language teaching stands to undergo a significant transformation. GPT's powerful language capabilities encompass semantic understanding, contextual comprehension, semantic coherence, and smooth question-answering.

This paper focuses on refining GPT's communication style by referencing daily scenarios to enhance interactive foreign language instruction. We examine the effectiveness of prompts and fine-tuning techniques to make conversations natural and cater content to learners' proficiency levels.

2. Limitations to Existing Approaches

Researchers currently use GPT's robust natural language generation in foreign language learning systems. Topsakal and Topsakal proposed a framework using AR, speech bots, and ChatGPT for language teaching tools targeting children. ChatGPT efficiently creates instructional content for the system [1]. Moreover, Xiao et al. implemented a GPT-based system for generating reading comprehension exercises based on input requirements. Evaluations show that GPT-generated content matches human-written articles in quality [2].

In a previous study, we proposed a GPT-based language learning system, offering interactive and personalized learning resources [3]. The system incorporates dialogue exercises resembling instant messaging applications, utilizing GPT to simulate authentic conversations across various topics. However, larger language models don't always ensure better adherence to user intents, they might produce unhelpful outputs or diverge from user expectations [4]. Compared to real-life conversation, interactions with GPT still have certain limitations, including three prominent issues:

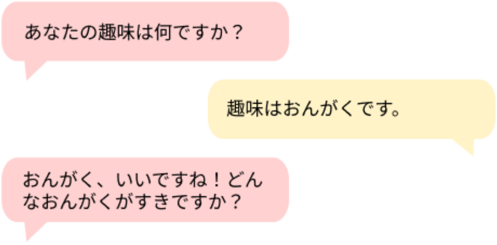
1) GPT responses often tend to be lengthy and intricate, whereas typical daily conversations that prioritize interaction over lengthy descriptions.
2) GPT tends to explain more than guide subsequent dialogues, lacking prompts like inquiries, leaving some users unsure how to continue the conversation.
3) Regardless of users' proficiency levels or language complexity, GPT responses to the same content do not significantly vary in length or difficulty.

3. Proposed Method

Based on the current limitations, we propose employing fine-tuning techniques and manually designing prompts to enhance GPT's ability to simulate daily conversations and match users' proficiency. We designed and compared three models: the original (GPT3.5), a fine-tuned version (GPT3.5-FT), and a pre-prompted version (GPT3.5-PR), all based on OpenAI's GPT-3.5-turbo.

3.1 Fine-tuning

Constructing datasets typically involves using extensive learner-GPT interactions for fine-tuning, but this is resource-intensive [5]. Due to cost and time constraints, we focused solely on 3-turn dialogues between GPT-learner-GPT when designing the dataset. The goal is to adapt GPT's 3rd-round responses based on the learner's 2nd-round input, considering aspects like length, content richness, vocabulary and grammar complexity, aligning with learner's proficiency level.



あなたの趣味は何ですか？

趣味はおんがくです。

おんがく、いいですね！どんなおんがくがすきですか？

Figure 1: 3-turn dialogue sample

We referenced Usami's Japanese Natural Conversation Corpus [6] and incorporated transcripts from real-life conversational scenarios to manually build dataset. It covers common topics in current foreign language oral practice, such as personal hobbies, daily life, sports, travel, and nature. The datasets were transformed into 3-turn dialogues: GPT's question, learner's answer, and GPT's response. During dataset curation, responses lacking substantial content (such as pure exclamations) and over lengthy statements without fine-tuning significance were filtered out, and guiding responses were added. This ensured the dataset focuses on responses that are relevant to the topic, providing meaningful and guiding contents.

3.2 Designing prompts

We constrain GPT's behavior by manually designing prompts. It's crucial to limit lengthy sentences and excessive questioning within prompts; otherwise, GPT may deviate from daily conversation style:

Now, you're a Japanese conversation practice *assistant. Please assess the user's language proficiency based on their vocabulary, grammar complexity, sentence length, and grammar errors. Respond at a level equivalent to the user's language proficiency. For instance, if the user's vocabulary and grammar are simple, adjust your response to match that simplicity; if they're more complex, respond accordingly. Additionally, in your replies, guide the user for the next part of the conversation by asking questions. Also, try to simulate daily conversations—avoid overly lengthy sentences or excessive consecutive questioning.*

4. Evaluation

Figure 2 shows the feedback from 15 international student users regarding their experiences with the three models, including the richness, length, readability, naturalness, difficulty of response and effectiveness of exercise. Among these, scores for length and difficulty should ideally be moderate, while for other dimensions, higher scores are preferred.

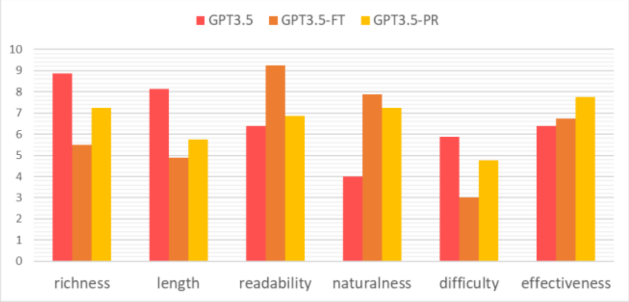


Figure 2: quality scores of the models in 6 dimensions

As depicted in the charts, both the richness of content in GPT3.5-FT and GPT3.5-PR have decreased due to limitations imposed on vocabulary difficulty and sentence length. However, they notably improved across five other aspects compared to GPT3.5. Both fine-tuning and prompts effectively condensed sentence length to an appropriate range, and scoring high in 'naturalness' for simulating natural conversation. GPT3.5-FT had better readability due to shorter sentences, but this limited complex language, relatively reducing exercise effectiveness. In contrast, GPT3.5-PR exhibited sentence lengths roughly between the GPT3.5 and GPT3.5-FT, thus displaying higher effectiveness. Nevertheless, without strict sentence structure and length limitations, occasional lengthy responses slightly impacted readability.

5. Discussion

We have demonstrated that fine-tuning can effectively improve GPT's conversational style, resulting in more natural and concise language expressions, which are more suitable for learners at intermediate to lower proficiency levels. However, as mentioned earlier, complex and lengthy expressions are restricted, reducing the effectiveness of practice for advanced learners. Therefore, increasing the vocabulary and grammar difficulty in high-level responses within the dataset can offer more advanced knowledge. Furthermore, although GPT can guide user conversations to a certain extent, limited dataset size restricts transitions across topics. Hence, it's necessary to expand the dataset to cover a broader range of topics.

On reflection, the authentic dialogue practice might be only suitable for learners at the relatively low to intermediate-level learners, but not for those at beginning or high proficiency levels. Beginners may face numerous unfamiliar words and complex sentences, impeding smooth communication. They may benefit from foundational vocabulary and grammar. Conversely, advanced learners might find this simplified conversational approach lacking in challenging content. For them, scenarios like school or business interviews, which demanding extensive narrative responses, could better serve their listening and speaking training requirements.

6. Conclusion

This paper proposed and validated the fine-tuning techniques and prompt design for optimizing GPT's conversational methods. This improvement creates natural dialogues suitable for learners of varying proficiency, providing tailored learning resources. The future lies in refining these models to provide personalized and effective language learning experiences for a wide range of learners.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP20H00622, JP23H03507, JP23K02469.

References

[1] Topsakal, O., & Topsakal, E., “Framework for a foreign language teaching software for children utilizing AR, voicebots and ChatGPT (Large Language Models)”, *Journal of Cognitive Systems*, 7(2), (2022).

[2] Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L., “Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications”, *Workshop on Innovative Use of NLP for Building Educational Applications*, (2023).

[3] Zhang, Z., Konomi, S., “GPTalkMate: Opportunities and challenges of large language models for foreign language education”, *Proceedings of the 22th IPSJ FIT*, (2023).

[4] Ouyang, L., et al., “Training language models to follow instructions with human feedback”, *Advances in Neural Information Processing Systems*, 35 (2022).

[5] Usami, M. (ed.), Building of a Japanese 1000 person natural conversation corpus for pragmatic analyses and its multilateral studies, and NINJAL Institute-based projects: Multiple Approaches to Analyzing the Communication of Japanese Language Learners, (2023).