Question Generation to Elicit Users' Food Preferences by Considering the Semantic Content

Jie Zeng and **Yukiko I. Nakano** and **Tatsuya Sakato**
Faculty of Science and Technology, Seikei University
{jie-zeng, y.nakano, sakato}@st.seikei.ac.jp

Abstract

To obtain a better understanding of user preferences in providing tailored services, dialogue systems have to generate semi-structured interviews that require flexible dialogue control while following a topic guide to accomplish the purpose of the interview. Toward this goal, this study proposes a semantics-aware GPT-3 fine-tuning model that generates interviews to acquire users' food preferences. The model was trained using dialogue history and semantic representation constructed from the communicative function and semantic content of the utterance. Using two baseline models: zero-shot ChatGPT and fine-tuned GPT-3, we conducted a user study for subjective evaluations alongside automatic objective evaluations. In the user study, in impression rating, the outputs of the proposed model were superior to those of baseline models and comparable to real human interviews in terms of eliciting the interviewees' food preferences.

1 Introduction

With interviews being used for various purposes, interview systems such as surveys (Johnston et al., 2013; Stent et al., 2006), job interviews (Inoue et al., 2020), and coaching (Hoque et al., 2013) have been developed. Interviews are categorized into three types: structured, semi-structured, and unstructured. In terms of flexibility, semi-structured interviews are between structured and unstructured. They are not completely planned but have a topic guide that needs to be covered. To build a dialogue system that can generate semi-structured interviews, flexible dialogue control must be provided while following the topic guide. To address the issues involved in generating semi-structured interviews, this study proposes an interview system to learn user food preferences.

Various dialogue control mechanisms have been studied in task-oriented dialogue systems to collect information from users, with the system responses are determined based on manually defined rules, POMDP (Young et al., 2010), deep learning (Chen et al., 2019), and reinforcement learning (Sankar and Ravi, 2019). However, these systems have less flexibility in dialogue control because the dialogue states are defined as a set of slot-value pairs that are limited to the task domain.

Research on open-domain non-task-oriented dialogue generation has contributed to the development of chitchat systems that can produce system responses for various topics. Initially, a simple sequence-to-sequence approach (Sordoni et al., 2015; Vinyals and Le, 2015; Serban et al., 2016; Li et al., 2016) was employed to generate a response. This approach has been improved to produce appropriate and meaningful responses, considering the dialogue context (Serban et al., 2017), and generate knowledge-grounded responses (Hedayatnia et al., 2020; Wu et al., 2020; Zhang et al., 2020a; Galetzka et al., 2021). More recently, ChatGPT (Ouyang et al., 2022) has demonstrated remarkable performance in generating rich and natural dialogues. However, these techniques have not yet been designed to generate dialogues for user model acquisition. Consequently, interview systems are required to generate responses that are aligned with the purpose of the interview.

To overcome the problems discussed above and generate useful questions in semi-structured interviews to elicit user food preferences, this study proposes a GPT-3 based model trained to generate responses with its semantic representation, which is constructed from the utterance's communicative function and semantic content. Semantic content refers to a structured sequence of labels for objects and their attributes. It is expected that using semantic content as part of the training targets would help constrain the generated responses towards eliciting the user food preferences.

The contributions of this study are as follows: 1) a semantic representation is proposed for sys-

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 190–196
September 11–15, 2023.

<Role (I/C)>-<message#>-<sentence#>	sentence	Communicative function	Semantic content
I-1-1	It's almost lunchtime, what do you eat for lunch?	Q-plan	[eat, [(Dish, ?)]]
U-2-1	Right.		
U-2-2	I like sandwiches.		
I-3-1	What do you like to have as sandwich ingredients?	Q-preference-positive	[like, [(Dish, sandwich, ingredient, ?)]]
U-4-1	I like tuna.		
I-5-1	Tuna is good on a sandwich.	Reply	[think, [(Dish, sandwich, ingredient, tuna)],[Evaluation, good]]
I-5-2	What do you often drink with your sandwich?	Q-habit	[drink, [(Drink,?, combine-with, sandwich)]]

Prompt and completion pairs in GPT-3 fine-tuning

```
HISTORY
SYSTEM: It's almost lunchtime, what do you eat for lunch?
USER: Right. I like sandwiches.
---
INFORMATION_FOR_SYSTEM_OUTPUT
COMMUNICATIVE_FUNCTION_LABEL: Q-preference-positive
SEMANTIC_CONTENT:{
VERB: like
OBJECT_TYPE: Dish
OBJECT_NAME: sandwich
OBJECT_ATTRIBUTE: ingredient
OBJECT_ATTRIBUTE_VALUE: ?
EVALUATION: None
}
->SYSTEM_OUTPUT: What do you like to have as
                 sandwich ingredients?
```

Figure 1: Overview of the proposed method. The left table shows an example dialogue between an interviewer (I) and a customer (C). The communicative function and semantic content of the interviewer's utterances are shown in the third and fourth columns, respectively. The right side shows the Prompt and Completion input for GPT-3 fine-tuning used to predict interview utterance I-3-1. The blue part indicates the prompt, and the green part indicates the completion. ***Bold italics*** indicate utterances or annotated values.

tem responses; 2) a response generation model is created for the interviewer's role; and 3) the effectiveness of the proposed method in eliciting user preferences is demonstrated through an evaluation experiment.

2 Corpus collection

To prepare the dataset used in this study, text-based dyad conversations were collected to interview participants regarding their food preferences. The participants were recruited through crowdsourcing. Each participant was assigned the role of either interviewer or interviewee and communicated using a chat system on a web browser. The interviewer was instructed to elicit the partner's preference for food, whereby they exchanged messages taking turns, for a minimum of 40 turns. Thus, a total of 118 Japanese dialogues were collected.

3 Method

To train a response generation model for the interviewer's role by considering the semantic representation of the interviewer's responses, we propose the method illustrated in Figure 1. First, the semantic representation of the interviewer's responses is presented, and subsequently, model training is explained.

3.1 Semantic representation of interviewer's responses

The semantic representation of an interviewer's utterance comprises the intention and meaning of the utterance. This representation can be exploited to train the dialogue generation model and direct the dialogue toward eliciting food preference information, as explained in detail below.

Communicative Function (CF): To specify the intention of the utterance, we refined the labels for self-disclosure and question types proposed in SWBD-DAMSL (Jurafsky, 1997) and Meguro et al. (2014), thereby defining 20 labels. The list is shown in the Appendix A.

Semantic Content (SC): The meaning of an utterance is described as a structured sequence of labels for verb and object features, such as `OBJECT_TYPE`, `OBJECT_NAME`, `OBJECT_ATTRIBUTE`, and `OBJECT_ATTRIBUTE_VALUE`.

Examples of semantic representation are shown in Figure 1. In utterance I-3-1, "What do you like to have as sandwich ingredients?" the communicative function is Q-preference-positive. The semantic content begins with the verb category. In this case, the verb is *like*. This is followed by object features `OBJECT_TYPE`: *Dish*, `OBJECT_NAME`: *sandwich*, `OBJECT_ATTRIBUTE`: *ingredient*, and `OBJECT_ATTRIBUTE_VALUE`: *?*. The *?* indicates that this value is missing. Thus, the semantic content of this utterance is expressed as [(Dish,sandwich,ingredient,?)]. Predefined values are used for the verbs and elements of `OBJECT_TYPE` and `OBJECT_ATTRIBUTE` for object features (see Appendix A). The details of the SC scheme were proposed in Zeng et al. (2022).

After annotating the CF and SC in the corpus collected in Section 2, we calculated the inter-coder reliability between two annotators. Cohen's Kappa value for CF was $\kappa = 0.72$ (substantial agreement), and the agreement ratio for verbs and object fea-

	BLEU-1	BLEU-2	BLUE-3	BLEU-4	BERTScore
ChatGPT	22.99	11.23	5.46	2.38	0.72
Seq2Seq	**25.11**	15.05	**8.11**	2.48	**0.75**
CF+SC	24.98	**15.23**	7.53	**2.71**	**0.75**

Table 1: Average BLEU scores and BERTScore on the test set. The best score for each column is highlighted in bold.

tures in SC between the two annotators was 0.72.

3.2 Interviewer response generation model

To create a response generation model, we fine-tuned OpenAI's GPT-3 (Brown et al., 2020), thereby referring to this proposed model as the CF+SC model. The model generates the completion part that follows the prompt. The formats for the prompt and completion are shown in Figure 1. Up to five messages preceding the prediction target interviewer's response were added to the prompt as dialogue history. The completion consisted of the annotated CF and SC (Section 3.1) and the interviewer's response sentence. The format of the completion part is indicated by green letters in Figure 1. When multiple sentences were included in the interviewer's message (turn), the last sentence, which usually contains the main claim, was used as the prediction target.

4 Experiment and evaluation

We evaluated the performance of the proposed CF+SC model using three comparison targets: the ground truth and two baselines.

Ground truth (GT): Actual utterances of the interviewers were used as the ground truth.

Fine-tuned GPT-3 (Seq2Seq): This simple fine-tuning model uses GPT-3. The model was trained without semantic representation (CF and SC) of the prediction targets. A sequence of preceding utterances was provided as prompt, and the model output was the interviewer's response.

Zero-shot ChatGPT (ChatGPT): OpenAI's ChatGPT model (reinforcement learning with human feedback and chat-optimized models (Ouyang et al., 2022)), specifically, gpt-3.5-turbo-0301, was adopted as the best general-purpose dialogue model. The zero-shot method was employed such that only the dialogue history and the system's role as an interviewer were provided as prompts [1]. The system was instructed to play the role of the interviewer and generate a response to elicit customer preferences by considering the context.

The temperature parameter for the three GPT-based models was set to 0. Thus, the generation was almost deterministic. While the CF+SC model generates both semantic representation and text of the response, we used the `SYSTEM_OUTPUT` part to extract the system response text. The CF+SC and Seq2Seq models generate a single sentence. Thus, in order to align the comparison conditions, when ChatGPT model generates multiple sentences, the last sentence, which tends to contain the main claim, was used in comparing with the ground truth.

The GPT-3 ("davinci" model) was fine-tuned using OpenAI's API. The model was trained for four epochs. The batch size was eight, and the learning rate was 0.05. The validation loss remained constant after epoch two. The number of instances used for training and validation were 1671 and 206, respectively.

4.1 Automatic evaluation

Table 1 shows the automatic objective evaluations, BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020b). The BLEU-2 and BLEU-4 scores for the CF+SC model are higher than the baselines. However, the CF+SC is slightly inferior to Seq2Seq in BLEU-1 and 3 and comparable to BERTScore. These automatic evaluation metrics measure word overlap or proximity in a word embedding space between the actual responses and model output. Therefore, it is known that such metrics do not properly evaluate appropriate responses that are not similar to GT and do not correlate well with human evaluations (Liu et al., 2016). To evaluate the validity of the generated output as an interviewer's response, we conducted a user study, as described in the next section.

[1]We also tested the few-shot ChatGPT, with prompts including two example responses accompanied by CF and SC, as shown in Figure 1. However, the model did not produce an output in the requested format (e.g., the `SYSTEM_OUTPUT` part was not produced).

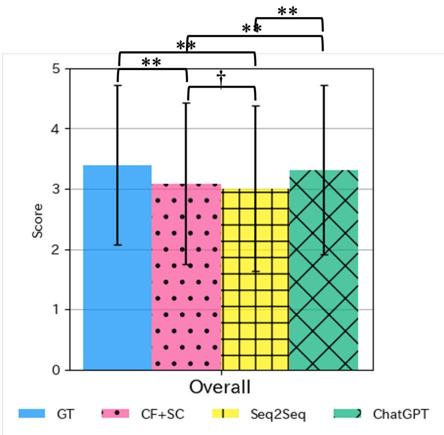

Figure 2: Overall impression evaluation result for interviewer response. The p-value was calculated using the Wilcoxon signed-rank test. ($\dagger$: $p < .1$ $**$: $p < .01$)

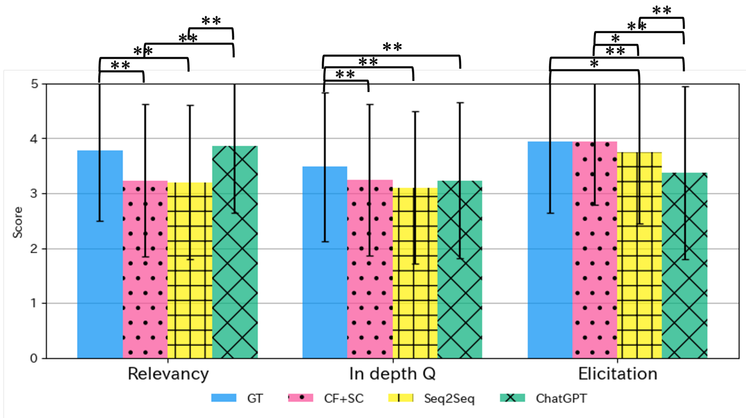

Figure 3: Impression evaluation regarding three detailed questions. The p-value was calculated using Tukey's HSD test. ($*$: $p < .05$ $**$: $p < .01$)

4.2 User study

For human evaluations, we conducted two user studies: 1) overall evaluation of responses from three models in addition to GT, and 2) ratings of one response from a single model.

1) Overall rating: A total of 460 experimental materials were created from the test set, each consisting of five preceding ground truth utterances as dialogue context, followed by a list of target responses from the four methods: GT, CF+SC (proposed model), Seq2Seq, and ChatGPT. The order of the target responses was randomized. The participants were instructed to rate the appropriateness of the interviewer's responses on a scale of 1 to 5 (a larger number is better). We recruited 30 participants through crowdsourcing and assigned 47 materials to each participant, including one to check for worker quality. Three ratings were collected for each material.

Figure 2 presents the results for the overall impression evaluation. GT and ChatGPT have similar scores which are significantly higher than those of the CF+SC and Seq2Seq models. The difference of CF+SC from Seq2Seq is marginally significant.

2) Ratings with clarified perspectives: In the second experiment, the following three questions were used to clarify the perspectives of the response ratings:

- **Relevancy:** Does the response fit the flow of the conversation?
- **In depth Q:** Does the response attempt to explore the interviewee's statements in depth?
- **Elicitation:** Does the response attempt to elicit information from the interviewee?

In this experiment, one target response was combined with five context utterances so that the subjects could not compare the responses from different methods. Participants were instructed to answer each of the three questions on a five-point Likert scale. We created 200 combinations of dialogue histories and the subsequent responses of each method. Thus, 800 materials were obtained, and 160 participants were recruited using crowdsourcing. Each worker was randomly assigned 21 materials (including one for worker quality check), and four participants evaluated each material.

The results are shown in Figure 3. Regarding relevancy, the performance of CF+SC is worse than that of ChatGPT and similar to that of Seq2Seq. For in depth Q, CF+SC is comparable to Seq2Seq and ChatGPT. Notably, in elicitation, CF+SC is equivalent to GT and superior to Seq2Seq and ChatGPT.

4.3 Discussion

In general, ChatGPT produced sentences that were as fluent and expressive as GT. Therefore, in the overall rating, the participants had a good impression of this model. The eloquence of ChatGPT may have led the participants to believe that the generated utterances fit the context (high relevancy). These results demonstrate the superior performance of ChatGPT as a general purpose dialogue model. Interestingly, ChatGPT performed the worst in the auto evaluation metric (Table 1), but the overall impression was the best. This confirms the low correlation between the subjective and objective evaluations discussed in Liu et al. (2016).

For asking in-depth questions (In depth Q), in all models, generated questions frequently included words used in the previous context. This is why we consider that subjects could not finding a clear difference between the three models in terms of the delving into the word that appeared in the context.

In elicitation, the proposed model (CF+SC) has a higher score than the other models. As shown in the Appendix, the CF+SC model is more likely to generate questions related to the objects and their attributes, indicating that CF+SC successfully considers semantic representation (Table 6 in the Appendix). Moreover, as shown in Table 7 in the Appendix, ChatGPT simply repeats the previous user's utterance in giving suggestions. These are not ideal responses for interviews. On the other hand, CF+SC asks questions that are not limited to the current context but covers broader aspects to actively elicit user preferences. We assume that these dialogue characteristics provide the subjects with the impression that the interviewer's response is an attempt to elicit user preferences. This suggests that semantic representation is important in training dialogue models for specific purposes.

5 Conclusions and future directions

This study proposed a response generation model aiming to extract user preferences for food. We trained the GPT-3 based model using a communicative function and semantic content. The results of the human impression evaluation experiment showed that the proposed model outperformed zero-shot ChatGPT and fine-tuned GPT-3 model, and comparable to real human interviews in terms of eliciting the interviewee's preferences.

One limitation of the current model is that it produces only a single sentence. In the future, this model should be improved to generate more complex responses using multiple sentences. Moreover, it is necessary to evaluate the model's performance in interactions with users, and examine whether the interview system is useful for understanding users.

Acknowledgements

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 and JST AIP Trilateral AI Research (PANORAMA project, grant no. JPMJCR20G6) and JSPS KAKENHI (grant number JP19H04159).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709.

Fabian Galetzka, Jewgeni Rose, David Schlangen, and Jens Lehmann. 2021. Space efficient context encoding for non-task-oriented dialogue generation with graph attention transformer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7028–7041.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421.

Mohammed (Ehsan) Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH: My automated conversation coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, page 697–706, New York, NY, USA. Association for Computing Machinery.

Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. Job interviewer android with elaborate follow-up question generation. pages 324–332. Association for Computing Machinery.

Michael Johnston, Patrick Ehlen, Frederick G Conrad, Michael F Schober, Christopher Antoun, Stefanie Fail, Andrew Hupp, Lucas Vickers, Huiying Yan, and Chan Zhang. 2013. Spoken dialog systems for automated survey interviewing. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 329–333.

Dan Jurafsky. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 110–119.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. 2014. Learning to control listening-oriented dialogue using partially observable markov decision processes. *ACM Transactions on Speech and Language Processing*, 10(4):1–20.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chinnadhurai Sankar and Sujith Ravi. 2019. Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. In *Proceedings of the 20th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–10, Stockholm, Sweden. Association for Computational Linguistics.

Iulian Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3288–3294.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3776–3783.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 196–205.

Amanda Stent, Svetlana Stenchikova, and Matthew Marge. 2006. Dialog systems for surveys: The rate-a-course system. In *Proceedings of the 2006 IEEE Spoken Language Technology Workshop*, pages 210–213.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online. Association for Computational Linguistics.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Jie Zeng, Tatsuya Sakato, and Yukiko Nakano. 2022. Semantic content prediction for generating interviewing dialogues to elicit users' food preferences. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 48–58, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Appendix

Table 2 shows the communicative function labels and Tables 3, 4, and 5 show the values used in the verbs, OBJECT_TYPE and OBJECT_ATTRIBUTE for the semantic content. Tables 6 and 7 present the dialogue history (-5 to -1) before the interviewer's response (GT) and the responses to CF+SC, Seq2Seq, and ChatGPT; I and C represent the interviewer and customer, respectively.

Information	SD-experience
SD-habit	SD-preference-positive
SD-preference-negative	SD-preference-neutral
SD-desire	SD-plan
SD-other	Q-information
Q-experience	Q-habit
Q-preference-positive	Q-preference-negative
Q-preference-neutral	Q-desire
Q-plan	Q-other
Proposal	Reply

Table 2: Communicative function labels (SD: Self-Disclosure, Q: Question)

Verb	Definition
like/!like eat/!eat recommend/ !recommend cook/!cook	
have/!have	Indicate that the user has a style or condition. Take Style, Condition for ObjectType.
think	e.g. "Pizza is the best food." → [think,[(Dish,Pizza)],[Evaluation,the best food]]
be	Describe universal knowledge. e.g. "Naengmyeon is Korean cuisine." → [be,[(Genre,Korean cuisine,type-of,naengmyeon)]]
other	Indicate a verb that does not fall into the above categories.

Table 3: Defined verb list. Notated as !+<*verb*> when defined for negative forms.

ObjectType	Definition	Example of ObjectName
Dish	Indicate dish.	curry and rice, hamburger
Ingredient	Indicate ingredient.	carrots, potatoes
Drink	Indicate drink.	juice, coffee
Food	Food or object rather than specific dishes or ingredients.	Do you have a favorite food? → [like,[(Food, ?)]]
Genre	Indicate genre or cuisine.	Japanese, Chinese
Texture	Indicate texture.	rich, creamy
Taste	Indicate flavor.	sweet, spicy
Style	Concepts related to food that correspond to ObjectName.	health, allergy
Condition	Indicate a condition that corresponds to ObjectName.	hunger, mood
Other	Food related items that do not fall under the above categories.	eating out, cup ramen

Table 4: Defined OBJECT_TYPE list

ObjectAttribute	Definition
ingredient	Represent the ingredient of the ObjectName.
type-of	Indicate the specific type of the ObjectName.
texture	Indicate the texture of the ObjectName.
taste	Indicate the taste of the ObjectName.
cooking-method+ eating-method	Indicate how to eat or make the ObjectName.
part-of	Indicate the part of the ObjectName.
combine-with	With or in combination with ObjectName.

Table 5: Defined OBJECT_ATTRIBUTE list

-5 C	麺類ではパスタが特に好きです。 (**I especially like pasta in noodles.**)
-4 I	あーいいですね。ガーリック系ですか、 クリーム系ですか？ (**Oh, that's good. Is it garlic-based or cream-based?**)
-3 C	どちらも大好物ですが、特にガーリック系は 美味しいですね。 (**I'm a big fan of both, but especially the garlic-based ones are delicious.**)
-2 I	お肉では鶏、牛、豚のどれが好きですか？ (**In terms of meat, which do you like best, chicken, beef, or pork?**)
-1 C	甲乙つけがたいですが、良く食べるのは鶏肉ですね。 さっぱりしていて好きです。 (**It's hard to choose, but I often eat chicken. I like its refreshing taste.**)
GT	ポトフと肉じゃがだったらどちらを選びますか？ (**If you had to choose between pot-au-feu and nikujaga, which would you choose?**)
CF+SC	COMMUNICATIVE_FUNCTION_LABEL: "Q-preference-positive" SEMANTIC_CONTENT: { VERB: "like" OBJECT_TYPE: "Ingredient" OBJECT_NAME: "鶏肉" (chicken meat) OBJECT_ATTRIBUTE: "作り方+食べ方" (cooking-method+ eating-method) OBJECT_ATTRIBUTE_VALUE: "?" EVALUATION: "None" } 鶏肉はどのような調理法で食べますか？ (**How do you cook chicken?**)
Seq2Seq	お魚はどうですか？ (**What about fish?**)
ChatGPT	では、焼き鳥や唐揚げ、どちらが好きですか？ (**Then, which do you prefer, yakitori or fried chicken?**)

Table 6: Example A

-5 C	和食を好んでいます。 (**I prefer Japanese food.**)
-4 I	和食の中で何がお好きでしょうか？ (**What do you like in Japanese food?**)
-3 C	魚が使われている料理が好きです。 (**I like dishes containing fish.**)
-2 I	魚は何の魚がお好きですか？ (**What kind of fish do you like?**)
-1 C	調理方法によりますが、ブリやアジが好きです。 (**It depends on how it is cooked, but I like yellowtail and horse mackerel.**)
GT	調理方法は何がお好きでしょうか。 (**What is your favorite cooking method?**)
CF+SC	COMMUNICATIVE_FUNCTION_LABEL: "Q-preference-positive" SEMANTIC_CONTENT: { VERB: "like" OBJECT_TYPE: "Ingredient" OBJECT_NAME: "魚" (fish) OBJECT_ATTRIBUTE: "種類" (type-of) OBJECT_ATTRIBUTE_VALUE: "?" EVALUATION: "None" } それではブリやアジ以外の魚はどんな魚が お好きですか？ (**So, what other types of fish do you like besides yellowtail and horse mackerel?**)
Seq2Seq	それとも魚介類はどうですか？ (**Or how about seafood?**)
ChatGPT	今後のメニューには、ブリやアジを使った和食を 取り入れてみてはいかがでしょうか？ (**How about including Japanese dishes that use yellowtail or horse mackerel in your future menu?**)

Table 7: Example B