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

Anonymous ACL submission

Abstract

To obtain better understanding of users' preferences in providing tailored services, dialogue systems need to generate semi-structured interviews, where flexible dialogue control is required while following a topic guide to accomplish the purpose of the interview. Toward this goal, this study proposes semantic-aware GPT-3 fine-tuning model that generates interviews to acquire users' food preferences. The model is trained using dialogue history and semantic representation, which is constructed from communicative function and semantic content of the utterance. In automatic-evaluation, the proposed method outperformed Zero-shot ChatGPT and fine-tuned GPT-3 without learning the utterance semantic representation. In impression rating in our user study, it was found that the proposed model was comparable to real human interviews in terms of eliciting the interviewee's food preferences.

1 Introduction

Interviews are used for various purposes, and 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. Semi-structured interviews are between structured and unstructured in terms of flexibility; they are not completely planned but use a topic guide that needs to be covered. To build a dialogue system that can generate semi-structured interviews, it is necessary to provide flexible dialogue control while following the topic guide to accomplish the purpose of the interview. By addressing these issues on generating semi-structured interviews, this study proposes an interview system that acquires a users' food preferences.

Various dialogue control mechanisms have been studied in task-oriented dialogue systems to collect information from users, where 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 limited to the task domain.

In contrast, research on generating open-domain non-task-oriented dialogues has contributed to developing chitchat systems that can produce system responses on various topics. Initially, a simple seq2seq approach (Sordoni et al., 2015; Vinyals and Le, 2015; Serban et al., 2016; Li et al., 2016) was employed in response generation, and improved to generate appropriate and meaningful responses by considering the dialogue context (Serban et al., 2017) and generate knowledge-grounded responses (Hedayatnia et al., 2020; Wu et al., 2020; Zhang et al., 2020; Galetzka et al., 2021). More recently, ChatGPT (Ouyang et al., 2022) has shown remarkable performance in generating rich and natural dialogue. However, these techniques have not been designed to generate dialogues to create a user model. Interview systems are required to generate responses oriented towards the interview's purpose.

To overcome the problems discussed above and generate in-depth questions in semi-structured interviews to elicit users' 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 is a structured sequence of labels for object(s) and their attributes. We expect that using semantic content as part of the training targets will contribute to constraining the generated responses towards eliciting the user's food preferences.

The contributions of this study are as follows: 1) proposing semantic representation of system responses, 2) creating a response generation model for the interviewer's role, and 3) showing the effec-

<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 in a sandwich?	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 during fine-tune on GPT-3

```
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 in a sandwich?
```

Figure 1: Overview of the proposed method. The left table shows an example dialogue between an interviewer (I) and a customer (C) as well as communicative function and semantic content. The right side shows Prompt and Completion input in GPT-3 training to predict interview utterance I-3-1. The blue part indicates prompt and the green part indicates completion. ***Bold italics*** indicate utterances or annotated values.

tiveness of the proposed method in eliciting user preferences by conducting an evaluation experiment.

2 Corpus collection

To prepare the dataset used in this study, we collected text-based dyad conversations to interview participants regarding their food preferences. We used crowdsourcing to recruit the participants. Each participant was assigned the role of either interviewer or interviewee and communicated using a chat system on a web browser. We instructed the interviewer to elicit the partner's preference for food. They exchanged messages by taking turns and were required to exchange more than 40 turns. We collected 118 dialogues in Japanese.

3 Method

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

3.1 Semantic representation of interviewer's responses

The semantic representation of interviewer's utterance comprises the intention and meaning of the utterance. We expect that exploiting this representation to train the dialogue generation model will contribute to directing dialogue toward eliciting food preference information. We have provided a detailed explanation below.

Communicative Function (CF): To specify the intention of the utterance, we defined 20 labels by refining the labels for self-disclosure and question types based on SWBD-DAMSL (Jurafsky, 1997) and Meguro et al. (2014). 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 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 of this utterance 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 value *?* indicates that this value is missing. Thus, the semantic content of this utterance is expressed as [(Dish,sandwich,ingredient,?)]]. We use predefined values for the verbs and the elements for `OBJECT_TYPE` and `OBJECT_ATTRIBUTE` in object features (see Appendix A).

We annotated CF and SC to the corpus collected in Secrtion 2. We calculated the inter-coder reliability between two annotators. The Cohen's Kappa value for CF was $\kappa = 0.72$ (substantial agreement), and the agreement ratio for verbs and object features in SC between two annotators was 0.72.

3.2 Interviewer response generation model

We created a response generation model by fine-tuning OpenAI's GPT-3 (Brown et al., 2020). 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 consists of the annotated CF and SC (Section 3.1) and the interviewer's response sentence. The format of the completion part is also indicated with green letters in Figure 1.

4 Experiment and evaluation

We evaluated the performance of the proposed model (hereafter referred to as CF+SC) with three comparison targets: ground truth and two baselines.

Ground truth (GT): We used actual utterances from the interviewers as the ground truth.

Fine-tuned GPT-3 (Seq2Seq): This is a simple fine-tuning model that uses GPT-3. We trained the model without semantic representation (CF and SC) of the prediction target. Thus, we provided a sequence of preceding utterances as a prompt, and the model output was the interviewer's response text.

Zero-shot ChatGPT (ChatGPT): We used OpenAI's ChatGPT model (reinforcement learning with human feedbacks and Chat-optimized models (Ouyang et al., 2022)), specifically, gpt-3.5-turbo-0301, as the best general-purpose dialogue model. The zero-shot method was employed, so that only the dialogue history and the system's role as an interviewer were given as the prompt. We instructed the system to play a role of an interviewer and generate a response to elicit customer preferences by considering the context.

The generation of the three GPT-based models was deterministic (temperature=0). While the CF+SC model generates both the semantic representation and text of the response, we used the `SYSTEM_OUTPUT` part to extract the system response text. When multiple sentences were included in the interviewer's message (turn), the last sentence, which usually contained the main claim, was used as the ground truth response. Similarly, the last sentence generated from the ChatGPT model was used in comparing with the ground truth.

We fine-tuned GPT-3("davinci" model) 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 2. The number of instances used for training and validation were 1671 and 206, respectively.

Model	BLEU-4 score
CF+SC	**2.71**
Seq2Seq	2.48
ChatGPT	2.38

Table 1: Average BLEU-4 scores on the test set

4.1 Automatic evaluation

We first calculated the BLEU-4 as an automatic objective evaluation. As shown in Table 1, the output sentences generated from the proposed model (CF+SC) were more similar to the ground truth than those of the other models.

4.2 User study

For human evaluation, we conducted two user studies:1) overall evaluation of responses from three models plus GT and 2) ratings of one response from a single model by asking more specific questions.

1) Overall rating: We created 460 experimental materials from the test set. Each material consisted of five preceding ground truth utterances as the 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 across materials. The participants were instructed to rate the responses on a scale of 1 to 5 (a larger number is better) regarding the appropriateness as interviewer's response. We recruited 30 participants through crowdsourcing and assigned each subject 47 materials, including one for worker quality check. Therefore, three ratings were collected for each material.

Figure 2 shows the results of the overall impression evaluation. GT and ChatGPT had similarly high scores, significantly higher than those of the CF+SC and Seq2Seq models. Next was CF+SC, and the difference from Seq2Seq was marginally significant.

2) Ratings with clarified perspectives: In the second experiment, we used the following three questions 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

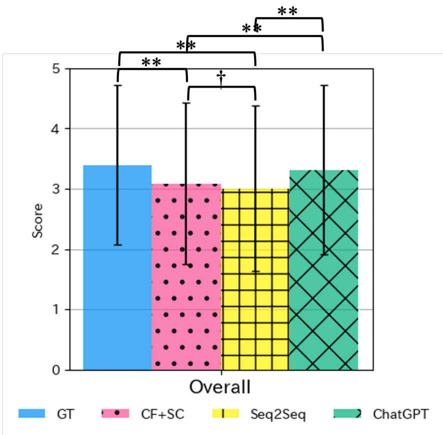

Figure 2: Overall impression evaluation result as interviewer response. We computed the p-value using a Wilcoxon signed-rank test. ($\dagger : p < .1$ $** : p < .01$)

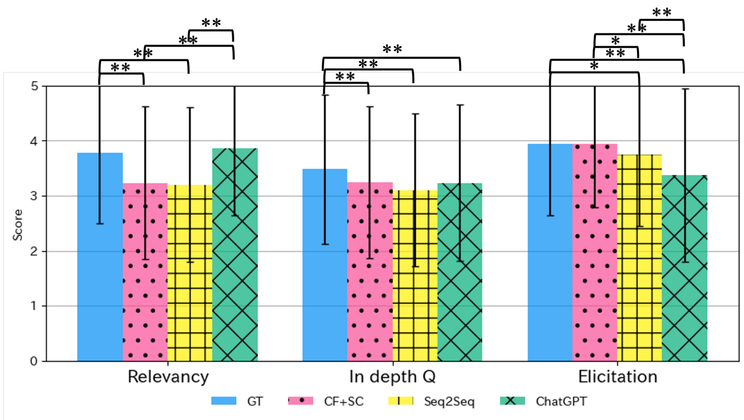

Figure 3: Impression evaluation regarding three detailed questions. We computed the p-value using a Tukey's HSD test. ($* : p < .05$ $** : p < .01$)

information from the interviewee?

In this experiment, one target response was combined with five context utterances, so that the subjects did not compare the responses from different methods. The participants were instructed to answer each of the three questions on 5 point Likert scale. We created 200 combinations of dialogue histories and the following responses for each method: Thus, we obtained a total of 800 materials and recruited 160 participants using crowdsourcing. Each worker was randomly assigned 21 materials (including one material for the worker quality check), and four participants evaluated each material.

The results are shown in Figure 3. Regarding Relevancy, CF+SC performed worse than ChatGPT and was nearly equal to Seq2Seq. CF+SC was comparable to Seq2Seq and ChatGPT for the in depth Q. Notably, for Elicitation, CF+SC was equivalent to GT and superior to Seq2Seq and ChatGPT.

4.3 Discussion

In general, ChatGPT produced sentences as fluent and expressive as GT. Therefore, in overall rating, the participants had a good impression to this model. This eloquence of ChatGPT may have led the subjects 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.

For asking in-depth questions (In depth Q), in most cases, all three models generate question that include the words used in the context utterances. Therefore, the three models were rated as comparable.

For Elicitation, the proposed model (CF+SC) had a higher score than the other generation models. As shown in Appendix, CF+SC model was more likely to generate questions related to the objects and their attributes indicating that the CF+SC successfully considered the semantic representation (Table 6). Moreover, CF+SC generated a more focused conversation than ChatGPT (Table 7). We assume that this characteristics of the dialogue gave subjects the impression that the interviewer's response was attempting to elicit user preferences. This result suggests that semantic representation is important in training dialogue models with specific purposes.

5 Conclusions and future directions

This study proposed a response generation model that aims to extract user preferences for food. We trained a GPT-3 based model using communication function and semantic content. The results of our user study showed that the proposed model was comparable to real human interviews in terms of eliciting the interviewees' preferences.

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

Acknowledgements

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. UbiComp '13, 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.

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.

Chinnadhurai Sankar and Sujith Ravi. 2019. Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. In *Proceedings of the 20th Annual SIGdial Meeting 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.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. 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.

A Appendix

Table 2 shows the communicative functioin labels, and Tables 3, 4 and 5 show the values used in verb, `OBJECT_TYPE`, and `OBJECT_ATTRIBUTE` for semantic content. Table 6 and 7 show the dialogue history (-5 to -1) before the interviewer's response (GT) and the responses by the three models: CF+SC, Seq2Seq, and ChatGPT. I and C represent 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 are a Korean cuisine." → [be,[(Genre,Korean cuisine,type-of,naengmyeon)]]
other	Indicates 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, Chienese
Texture	Indicate texture.	rich, creamy
Taste	Indicate flavor.	sweet, spicy
Style	Concepts related to food that corresponds 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	Represents 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 are 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 using yellowtail or horse mackerel in your future menu?**)

Table 7: Example B