Detecting Individual Decision-Making Dialogues in Conversation

Weiwen Su[1] Naoki Yoshinaga[2] Masashi Toyoda[2] Zihan Wang[1] Yuhan Zhou[1]
[1]The University of Tokyo [2]Institute of Industrial Science, The University of Tokyo
{su-w,ynaga,toyoda,zwang,yzhou}@tkl.iis.u-tokyo.ac.jp

Abstract

Decision-making is an essential part of our daily lives, especially in dialogue. It involves group decision-making, where we strive to reach a consensus with others, or individual decision-making primarily based on our independent thinking. Collecting decision-making data helps us analyze our daily behaviors and engage in self-reflection. In this study, aiming to detect individual decision-making dialogues in conversation automatically, we annotate the decision-seeking (*e.g.*, "Would you like to form a band with me?") and decision-making utterances in a dialogue dataset. We then investigate the LLMs' ability to detect individual decision-making and conduct an error analysis to analyze the mistakes in the detection processes.

1 Introduction

We engage in decision-making during daily dialogue, as illustrated in Figure 1, whether individually or as a group. Decision-making is a fundamental cognitive process of human behavior [1] and reflects one's thinking. By analyzing our decision-making in daily dialogues, we can gain a better understanding of our behaviors, which may also foster self-reflection. However, detecting decision-making processes within dialogue remains a challenging task.

In the early year, Fernandez et al. [2] explored the task of detecting group decision-making in conversation using a meeting corpus and proposed a decision dialogue acts (DDAs) class set. Recently, encouraged by the strong language understanding ability of pre-trained language models (PLMs), Karan et al. [3] revisited the task and discovered that the models sometimes depend on topic-specific words for detection. However, even with the PLMs, the performance of the task remains quite a space to be improved. Moreover, the detection of individual decision-making that often happens in daily dialogue and is related to one's personal choice has not yet been thoroughly investigated, which remains an obstacle to understanding one's behaviors through decision-making in conversation. In this study, we explore detecting individual decision-making dialogues in conversation by building a dataset for evaluation and testing current models on the dataset. Since spontaneous individual decision-making (suddenly announcing a decision without being requested) often does not offer an apparent trigger point, we focus on the individual decision-making upon request, meaning that there is a decision-seeking utterance explicitly or implicitly requests the interlocutor to make a decision, and then the interlocutor replies with a decision-making utterance. Besides, there might be a discussion between decision-seeking and decision-making that makes the detection more difficult considering the possible change of decision. To approach the goal, we first choose the TvShowGuess [4] dataset containing some daily drama scripts as a source dialogue dataset. Then we manually annotate the pairs of decision-seeking (Utterance 1 in Figure 1) and decision-making utterances (Utterance 4 in Figure 1) with their possible discussion process (Utterance 2 and 3 in Figure 1) in each dialogue. Additionally, we annotate the importance level (how much the decision would influence the decision-maker or other people) of the decisions in the dataset to assess the number of important

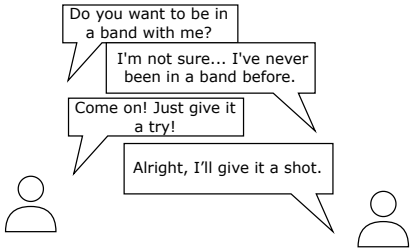

Figure 1 An example of individual decision-making dialogue upon request.

decision-making instances present.

Subsequently, we conduct experiments to investigate the current models' ability to detect individual decision-making dialogues in conversation and explore the possible methods to enhance the ability. Therefore, We evaluate the dataset with open-sourced and closed-sourced LLMs using several prompting methods or masking background information. Finally, we conduct an error analysis to discover possible reasons for the mistakes during the detection.

2 Related Work

In this section, we introduce the research on detecting decision-making in conversation. Hsueh et al. [5] initiated to explore the automatic detection of decision-making sub-dialogues in conversation (meeting corpus) by classifying the decision-related utterances using various features (*e.g.*, dialogue acts, prosodic features). Fernandez et al. [2] then proposed a set of decision dialogue acts and used a support vector machine (SVM) to classify the acts for detecting decision-making. Bui et al. [6] tried to improve the performance using hierarchical graphical models.

Recently, with the emergence of the pre-trained language models, Karan et al. [3] revisited the decision-making sub-dialogues detection task and tested the performance of BERT [7] model while revealing that sometimes the models depend more on the topic related words than the words indicating decision-making to do detection. Considering the emerging e-mail interaction in recent years, Karan et al. [8] proposed a large-organization email dataset for analyzing the dialogue acts within the decision-making.

The previous research mainly targets the decision-making dialogues in the business area using a meeting corpus. Besides, they show that the detection of decision-making, even detecting only the sub-dialogue rather than specific utterances, remains difficult for pre-trained language models to achieve. In this study, we focus on individual decision-making in open-domain dialogue where the topic is more related to daily life and the decisions are primarily based on independent thinking.

3 Decision-Making Dialogues

Due to the lack of datasets for decision-making detection in open-domain dialogue, this study builds a dataset to evaluate LLM-based baselines and demonstrate the complexity of the task. This section describes our approach to creating a small-scale dataset for detecting required individual decision-making in conversation.

In this study, we focus on the decision-making requested by the interlocutor because it has a relatively apparent trigger point compared with spontaneous decision-making where the speaker often suddenly announces a decision. We consider the dataset to involve some dialogues including decision-making pairs including:

Decision-seeking utterance An utterance where the speaker explicitly or implicitly requests the interlocutor to make a decision.

Corresponding decision-making utterance An utterance where the interlocutor makes a decision requested in the decision-seeking utterance. There might be decision-making utterances expressing tentative decisions, but only the final one is paired with the decision-seeking utterance.

To create such a dataset, we first choose a source dialogue dataset, the TVShowGuess dataset [4], as the starting point. It contains some daily drama scripts, which we can use as a multi-party open-domain dialogue dataset. Then three graduate students (the first, fourth, and fifth authors) annotate the dataset, concretely, to label the decision-seeking utterances and corresponding decision-making utterances in each dialogue.

To better analyze the decision-making process, we also ask the annotators to label the utterances between the decision-seeking and decision-making utterances. We consider the exchange of opinion and information would influence the final decision. Therefore, we formulate the categories of the middle utterances as "tentative decision", "asking for opinion", "providing opinion", "asking for information", "providing information", "agreement or disagreement", and "other utterances."

In addition, because there are many trivial decision-making (*e.g.*, "Do you want to have lunch together?") in daily dialogue, we ask the annotators to label the importance level (subjectively, how influential a decision is to the decision-maker or others) of the decision requested in the decision-seeking utterance to see how many important decisions exist in the dataset. The labels of importance level are "3, Highly influential or having a long-term impact on decision-makers or others.", "2, Influential to decision-maker or other people", and "1, Not influential".

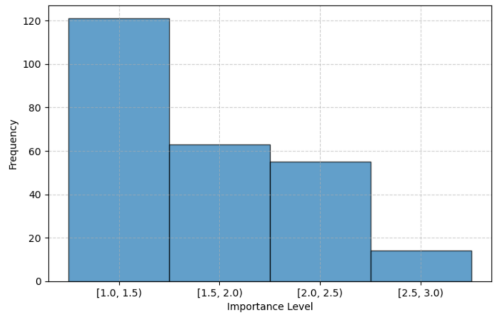

Figure 2 Distribution of the importance level of the decisions.

Table 1 Statistics of the dataset.

Dialogues	130
Utterances per dialogue	49.8
Decision-making pairs per dialogue	1.95

To ensure the precision of decision-making pairs in the dataset, we only pick the dialogues with decision-making pairs labeled by at least two annotators. To ensure the decision-making pairs in these dialogues are all labeled, we double-checked the decision-making pairs labeled by only one annotator in these dialogues and kept the ones that passed the double-check. As a result, we show the statistics of the dataset in Table 1. We take an average of the importance level from annotators for each decision and show the distribution in Figure 2. The distribution shows that nearly half of the required individual decision-making is influential to some extent.

4 Detecting Decision-Making

In this section, we evaluate two LLMs on the created dataset and conduct a case study to analyze the possible reasons for the mistakes in the detection. We start with the settings of the experiments, and then introduce the results and case study.

4.1 Settings

As for the task setting, the models need to detect all the decision-making pairs in each dialogue. Given a dialogue D with each utterance assigned a unique number u_i ($i = 1, 2, 3, 4, \ldots$), we ask the model to return each pair of decision-seeking and decision-making utterance as $[u_s, u_m]$, where u_s and u_m represent the utterance number of the two utterances.

To show the models' ability to detect decision-making pairs, we conduct an automatic evaluation. In addition, we also want to know which part of the models would make more mistakes. Therefore, we also evaluate the performance of detecting decision-seeking and decision-making utterances separately.

Models We evaluate one open-sourced model and one closed-sourced model on the dataset, the Llama3.1-8B-Instruction model [1)] and the GPT-4o model. As the input settings for the models, we test the following settings:

Zero-shot The models are given the description of the task and the definitions of decision-seeking utterance, decision-making utterance, and corresponding decision-making utterance as shown in the Appendix Table 4.

One-shot The models are additionally given one example as shown in the Appendix Table 6.

Chain-of-Thought (CoT) Considering the dependency between decision-seeking and decision-making utterances, we applied CoT [9] prompting method that often enhances the LLMs' reasoning ability as shown in the Appendix Table 5.

Anonymous To check whether the background knowledge of the speakers (*e.g.*, the relationships between the speakers or their characteristics) will influence the performance, based on the zero-shot setting, we replace the place, person, and organization name with placeholders (*e.g.*, *Joey* → *person_1*).

Metrics As for the metrics, we choose the precision ((**P**), recall ((**R**), and f1-score ((**F1**) to automatically evaluate the performance of the models. Concretely, we calculate the three metrics for the decision-making pairs, the decision-seeking utterances, and the decision-making utterances separately, between the gold labels and the predicted labels. For the decision-seeking utterances, we name the metrics $\mathbf{P}_s$, $\mathbf{R}_s$, and $\mathbf{F1}_s$ ($\mathbf{P}_m$, $\mathbf{R}_m$, and $\mathbf{F1}_m$ for decision-making utterances).

4.2 Main Results

Table 2 shows the automatic evaluation results of detecting individual decision-making dialogues in conversation. From the results of the GPT-4o model, we can observe that the zero-shot setting achieves the best overall performance and CoT setting achieves the worst overall performance. The reason may be that the GPT-4o model has better thinking steps than our prompt. As for the one-shot setting,

1) https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Table 2 Automatic evaluation results for GPT-4o and the Llama3.1-8B-Instruction model, including the results of overall detection, decision-seeking detection, and decision-making detection.

Settings	P	R	F1	P_s	R_s	$F1_s$	P_m	R_m	$F1_m$
GPT-4o									
Zero-shot	**0.436**	0.451	**0.443**	**0.562**	0.580	**0.571**	**0.540**	**0.540**	**0.540**
One-shot	0.416	0.451	0.433	0.514	0.558	0.535	0.511	0.535	0.523
CoT	0.403	0.425	0.414	0.525	0.553	0.539	0.481	0.491	0.486
Anonymous	0.403	**0.460**	0.430	0.519	**0.593**	0.554	0.510	**0.540**	0.525
Llama3.1-8B-Instruction									
Zero shot	0.073	0.248	0.113	0.141	0.460	0.216	0.147	0.434	0.219
One-shot	0.065	0.274	0.106	0.126	**0.491**	0.200	0.119	0.425	0.185
CoT	0.096	0.283	0.143	0.163	0.456	0.240	0.167	**0.456**	0.244
Anonymous	**0.106**	**0.317**	**0.159**	**0.177**	0.489	**0.260**	**0.177**	0.448	**0.254**

the performance does not exceed the zero-shot setting, the reason might be that the one-shot example fails to pinpoint the specific weaknesses or areas of difficulty in the model's detection capabilities.

From the results of the Llama3.1-8B-Instruction model, we can observe that the CoT setting and anonymous setting achieve better overall performance than the zero-shot setting. Especially, the anonymous setting performs the best, the reason may be that the background knowledge of the entities distracts the attention of the model. The CoT prompt still benefits the Llama3.1-8B-Instruction model for detecting decision-making pairs. Moreover, while the Recall of the detection is not particularly poor, the Precision is significantly low, indicating that the model struggles to distinguish incorrect samples. Therefore, introducing more counterexamples into the instruction might help improve the model's performance.

4.3 Error Analysis

We conduct an error analysis for the GPT-4o zero-shot setting to analyze the mistakes within the detection. We first count the samples of decision-making pairs with or without discussion (middle utterances between the decision-making pair) in the dataset and count that in the missed gold decision-making pairs. We find that 51.9% of the decision-making pairs with discussion are missed and 47.7% of the decision-making pairs without discussion are missed. That shows the decision-making pairs with discussion is more difficult for the model to detect.

To gain a deeper understanding of the errors, we randomly pick 8 samples of missed gold decision-making pairs and 8 samples of wrongly predicted decision-making pairs. Our observation of the samples shows that the implicit decision-seeking (*e.g.*, through suggestion or subtly bringing up) and decision-seeking utterance mixed with the response to the previous utterance cause some difficulties. In addition, We observe that the model occasionally misinterprets some decision-making utterances that are not responses to decision-seeking utterances as if they were.

Table 3 An example of a missed pair.

```
A: Yeah, call it whatever you want, I get minimum
wage. Yeah, anyway, I was wondering if you could
help me out with something, I was....
B: Yes.
```

Therefore, enhancing the models' ability to understand implicit decision-seeking utterances and correctly recognize whether an utterance is responding to the decision-seeking utterance may be a direction to improve the performance. Meanwhile, detecting the middle utterances may support the detection of the decision-making pairs.

5 Conclusions

In this study, we propose to detect individual decision-making dialogues in conversation by detecting the pair of decision-seeking and decision-making utterances. We create a dataset for detecting individual decision-making dialogues by annotating the pairs of decision-seeking and decision-making utterances and the discussion process. We then evaluate open-sourced and closed-sourced models on the dataset with an automatic evaluation. The results show that the current models need a well-designed method to accomplish this task. Besides, our error analysis shows that implicit decision-seeking utterances, exchanging opinions during discussion, and decision-making utterances that do not respond to the decision-seeking utterances pose some challenges for the task.

Acknowledgement

This work was partially supported by JST, CREST Grant Number JPMJCR19A4, JSPS KAKENHI Grant Number JP21H03494, and JSPS KAKENHI Grant Number JP21H03445

References

[1] Vassilios N. Christopoulos, Kristen N. Andersen, and Richard A. Andersen. Chapter 8 - extinction as a deficit of the decision-making circuitry in the posterior parietal cortex. In Giuseppe Vallar and H. Branch Coslett, editors, **The Parietal Lobe**, Vol. 151 of **Handbook of Clinical Neurology**, pp. 163–182. Elsevier, 2018.

[2] Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. Modelling and detecting decisions in multi-party dialogue. In David Schlangen and Beth Ann Hockey, editors, **Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue**, pp. 156–163, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[3] Vanja Mladen Karan, Prashant Khare, Patrick Healey, and Matthew Purver. Mitigating topic bias when detecting decisions in dialogue. In Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li, editors, **Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 542–547, Singapore and Online, July 2021. Association for Computational Linguistics.

[4] Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. Tvshowguess: Character comprehension in stories as speaker guessing. **arXiv preprint arXiv:2204.07721**, 2022.

[5] Pei-Yun Hsueh and Johanna D. Moore. What decisions have you made?: Automatic decision detection in meeting conversations. In Candace Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, **Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference**, pp. 25–32, Rochester, New York, April 2007. Association for Computational Linguistics.

[6] Trung H. Bui and Stanley Peters. Decision detection using hierarchical graphical models. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre, editors, **Proceedings of the ACL 2010 Conference Short Papers**, pp. 307–312, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, June 2019.

[8] Vanja Mladen Karan, Prashant Khare, Ravi Shekhar, Stephen McQuistin, Ignacio Castro, Gareth Tyson, Colin Perkins, Patrick Healey, and Matthew Purver. LEDA: a large-organization email-based decision-dialogue-act analysis dataset. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 6080–6089, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. **ArXiv**, Vol. abs/2205.11916, , 2022.

Appendix

Table 4 Prompt for zero-shot decision-making pair detection.

You are an expert in linguistics. Please analyze the given dialogue and identify all pairs of decision-seeking utterances and their corresponding decision-making utterances related to personal decision-making. 1. A decision-seeking utterance is an utterance where the speaker explicitly or implicitly encourages the interlocutor to make a decision. 2. A decision-making utterance is an utterance where the speaker makes a decision. 3. A corresponding decision-making utterance is one that responds to the decision requested in the decision-seeking utterance. Personal decision-making refers to decisions about an individual's own actions, choices, or preferences. For each decision-seeking utterance, there might be several tentative decision-making utterances. Pair it only with the final decision-making utterance. Each utterance in the dialogue is assigned a unique number. Please return only the numbers, formatted as a list of pairs: [(decision-seeking number, decision-making number)]. Only a list! No additional text!

Table 5 Prompt for CoT decision-making pair detection.

You are an expert in linguistics. Please analyze the given dialogue and identify all pairs of decision-seeking utterances and their corresponding decision-making utterances related to personal decision-making. 1. A decision-seeking utterance is an utterance where the speaker explicitly or implicitly encourages the interlocutor to make a decision. 2. A decision-making utterance is an utterance where the speaker makes a decision. 3. A corresponding decision-making utterance is one that responds to the decision requested in the decision-seeking utterance. Personal decision-making refers to decisions about an individual's own actions, choices, or preferences. For each decision-seeking utterance, there might be several tentative decision-making utterances. Pair it only with the final decision-making utterance. You should first find a decision-seeking utterance and then find the corresponding decision-making utterance answering that. Do it step by step. Each utterance in the dialogue is assigned a unique number. Please return only the numbers, formatted as a list of pairs: [(decision-seeking number, decision-making number)]. Only a list! No additional text!

Table 6 Prompt for one-shot decision-making pair detection.

You are an expert in linguistics. Please analyze the given dialogue and identify all pairs of decision-seeking utterances and their corresponding decision-making utterances related to personal decision-making. 1. A decision-seeking utterance is an utterance where the speaker explicitly or implicitly encourages the interlocutor to make a decision. 2. A decision-making utterance is an utterance where the speaker makes a decision. 3. A corresponding decision-making utterance is one that responds to the decision requested in the decision-seeking utterance. Personal decision-making refers to decisions about an individual's own actions, choices, or preferences. For each decision-seeking utterance, there might be several tentative decision-making utterances. Pair it only with the final decision-making utterance. Each utterance in the dialogue is assigned a unique number. Please return only the numbers, formatted as a list of pairs: [(decision-seeking number, decision-making number)]. Only a list! No additional text! **********Example********** Dialogue: (1) Speaker_A: Do you want to be in a band with me? (2) Speaker_B: I'm not sure... I've never been in a band before. (3) Speaker_A: Come on! Just give it a try! (4) Speaker_B: Alright, I' ll give it a shot. Your Return: [(1,4)]