Constructing a Multi-Party Conversational Corpus Focusing on Interlocutor Relationships

Taro Tsuda[1], **Sanae Yamashita**[1],
Koji Inoue[2], **Tatsuya Kawahara**[2], **Ryuichiro Higashinaka**[1]
[1]Graduate School of Informatics, Nagoya University
[2]Graduate School of Informatics, Kyoto University
{tsuda.taro.d7@s.mail., yamashita.sanae.w7@s.mail., higashinaka@i.}nagoya-u.ac.jp
{inoue.koji.3x@, kawahara@i.}kyoto-u.ac.jp

Abstract

Despite recent breakthroughs in dialogue systems, few systems are designed to handle multi-party conversations involving three or more interlocutors. One reason for this limitation is the lack of suitable corpora. To address this gap, we constructed the Multi-Relational Multi-Party Chat Corpus (MRMP), which consists of Japanese text-based chats comprising two types of conversation: first-time-meeting dialogues, in which all participants are meeting for the first time, and family-included dialogues, which involve both participants who are family members and participants meeting for the first time. In this paper, we present analyses of the corpus to examine how the relationships of interlocutors influence conversations.

1 Introduction

Recent advances in large language models (LLMs) have led to a significant improvement in the performance of dialogue systems (Brown et al., 2020; Shuster et al., 2022; OpenAI, 2023; Iizuka et al., 2023). The majority of these systems primarily engage in two-party interactions between a single user and the system. However, in real-world settings, conversations frequently involve three or more participants. Given the potential societal applications of dialogue systems, it is essential to develop systems capable of handling multi-party dialogues, but at present, the availability of dialogue resources for constructing such systems remains limited.

In this study, we construct the Multi-Relational Multi-Party Chat Corpus (MRMP), a Japanese text-based multi-party dialogue corpus. This corpus focuses on the relationships of interlocutors, which play a crucial role in multi-party dialogues involving three or more participants. Specifically, we define two types of interlocutor relationships—first-time and family—and incorporate interlocutors with these relationships into the dialogues. The dataset comprises two types of dialogue: (1) dialogues among three first-time interlocutors (first-time-meeting dialogues) and (2) dialogues involving two family members and one first-time interlocutor (family-included dialogues). In this paper, we analyze the constructed corpus to elucidate the impact of the relationships of interlocutors on dialogue, which can be beneficial for the development of multi-party dialogue systems. The key findings are as follows:

- In multi-party dialogues, participants tend to experience higher satisfaction levels when conversing with others in the same category (first-time/family).
- In first-time-meeting dialogues, participants actively express agreement and empathy to build relationships. In contrast, in family-included dialogues, the individual interlocutor tends to ask questions directed at the family interlocutors, and mentions (e.g., @[handle name]) between the individual interlocutor and the family interlocutor are frequently used.
- Humans can predict first-time relationships with approximately 60% accuracy, whereas family relationships are predicted with only 20% accuracy. Compared to humans, GPT-4o demonstrates limited capability in distinguishing first-time and family relationships from acquaintances.

To promote research on multi-party dialogue systems, we have made the MRMP corpus publicly available.[1]

2 Related Work

Existing multi-party dialogue corpora encompass a variety of settings (Mahajan and Shaikh, 2021),

[1]https://github.com/nu-dialogue/multi-relational-multi-party-chat-corpus

Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue, September 3–5, 2025, Bielefeld, Germany.

including daily conversations (Den and Enomoto, 2007; Koiso et al., 2022), meetings (Janin et al., 2003; Carletta et al., 2005), dialogues in radio and television broadcasts (Hsu et al., 2018; Poria et al., 2019; Chen et al., 2020; Zhu et al., 2021), and text-based chats (Lowe et al., 2015). However, many of these corpora do not explicitly specify the relationships of interlocutors, and they do not necessarily contain a sufficient amount of dialogue data involving interlocutors with specific relationships. In this study, we construct a multi-party dialogue corpus in which the relationships of interlocutors are systematically controlled to include either first-time interlocutors or family members.

Research on the relationships of interlocutors in multi-party dialogues includes works by Takanashi and Isahara (2002), who collected dialogues involving friends and first-time interlocutors, and Tsubokura et al. (2022), who gathered dialogues between a dialogue robot and parent-child pairs. These studies, which were limited in scale, aimed primarily at conversation analysis. In contrast, the present study collects a large number of dialogues, aiming for data-driven research.

3 Multi-Relational Multi-Party Chat Corpus

We constructed the Multi-Relational Multi-Party Chat Corpus (MRMP) in the following sequence: recruitment of interlocutors, administration of a pre-experiment questionnaire, collection of dialogue data, and administration of a post-experiment questionnaire. In this section, we provide a detailed description of each step and present statistical information on the collected data. The data collection process underwent an ethical review by our institution, including an assessment of the collection procedures and data storage methods.

3.1 Recruitment of Interlocutors

We recruited 66 interlocutors through crowdsourcing. Each interlocutor participated either individually or in pairs. In the latter case, pairs were restricted to family members within the second degree of kinship. The gender ratio was approximately 2:8 (male to female), and the age distribution was roughly 20–30% for each of the 30s, 40s, and 50s age groups.

3.2 Administration of Pre-Experiment Questionnaire

Prior to collecting dialogue data, each interlocutor completed a pre-survey regarding their personality traits, personal attributes, and experience with text-based chat. Additionally, each interlocutor was asked to create ten concise sentences representing themselves as a persona. The details of the pre-survey items and persona creation process follow the methodology outlined in previous research (Yamashita et al., 2023).

3.3 Collection of Dialogue Data

The interlocutors engaged in text-based chat sessions in groups of three. The dialogues were categorized into two types based on the relationships among the interlocutors. The first type was first-time-meeting dialogues. In these dialogues, all three interlocutors were meeting for the first time. Here, the interlocutors are referred to as first-time interlocutors. The second type was family-included dialogues. In these dialogues, each group consisted of a pair of family members and a single interlocutor participating independently. The pair of family members are referred to as family interlocutors, and the single interlocutor participating independently is referred to as an individual interlocutor. The individual interlocutor had no prior acquaintance with either family interlocutor.

The dialogues were conducted with a target of at least 100 utterances per session, with each utterance limited to a maximum of 50 characters. Interlocutors could enter text at any time without predefined turn-taking. Each interlocutor was assigned a unique handle name, and when addressing a specific interlocutor, they used the mention function of the text chat tool in the format @[handle name]. To prevent imbalances in interlocutor pairings, no trio of interlocutors was allowed to engage in more than ten dialogues together. Additionally, each interlocutor was restricted to participating in no more than 50 dialogues. The topic of conversation was unrestricted.

After each dialogue session, the participants completed a post-dialogue questionnaire in which they rated the informativeness and clarity of the other two interlocutors' utterances, as well as their perceived familiarity with the other two interlocutors. Additionally, they rated their own level of interest, proactiveness, and overall satisfaction. Each item was rated on a five-point scale.

	Multi-Relational Multi-Party Chat Corpus		CEJC
	First-time-meeting dialogues	Family-included dialogues	
No. of dialogues	500	500	297
No. of participants	40	26 (6 family pairs: 4 parent-child pairs, 2 married couples; 12 family interlocutors, 14 first-time interlocutors)	677
No. of utterances per dialogue	104.7 (100–168)	105.0 (100–152)	1221.8
No. of utterances	52,367	52,494	362,888
No. of characters per utterance	10.8 (1–53)	12.3 (1–62)	6.7
Vocabulary	12,295	13,885	44,069
No. of tokens	329,406	372,399	1,264,683
Type-Token ratio	0.037	0.037	0.034

Table 1: Basic statistics of Multi-Relational Multi-Party Chat Corpus and CEJC (Koiso et al., 2022) for reference. The CEJC statistics are aggregated for dialogues involving three or more interlocutors, and utterance counts are based on long utterance-units (Den et al., 2010).

3.4 Administration of Post-Experiment Questionnaire

To confirm the overall impression of the dialogues, each interlocutor provided a free-form response describing the dialogue that made the strongest impression on them after completing all their dialogues.

3.5 Statistical Information

Table 1 presents the basic statistics of the MRMP corpus. For comparison, the statistics of an existing multi-party dialogue corpus, the Corpus of Everyday Japanese Conversation (CEJC) (Koiso et al., 2022), are also provided. Note that CEJC consists of spoken dialogues, whereas MRMP comprises text-based chat dialogues.

The MRMP corpus we constructed comprises 500 first-time-meeting dialogues and 500 family-included dialogues. The first-time-meeting dialogues involved 40 interlocutors, while the family-included dialogues included 12 family members (six pairs) and 14 first-time interlocutors. The number of utterances per dialogue did not differ significantly between the two dialogue types. The average number of characters per utterance was slightly above 10, indicating that most utterances were relatively short. The type-token ratio for both dialogue types was comparable to that of CEJC, which includes a wide variety of conversational contexts and vocabulary, suggesting that a diverse range of vocabulary was used.

4 Comparison of Dialogue Types

We conducted a comparison between first-time-meeting dialogues and family-included dialogues to clarify their respective characteristics. Specifically, we analyzed the results of the post-dialogue questionnaire, the usage of mentions, and the linguistic expressions specific to each dialogue type.

	First-time meeting dialogues	Family-included dialogues		
	First-time	Avg.	Individual	Family
Informativeness	**4.50****	4.37	4.23	4.45††
Comprehension	**4.67****	4.49	4.37	4.54††
Familiarity	**4.56****	4.51	4.43	4.55††
Interest	4.44*	4.42	4.17	**4.54**††
Proactiveness	4.44	4.45	4.26	**4.54**††
Satisfaction	4.47	4.46	4.26	**4.56**††

Table 2: Results of post-dialogue questionnaire (five-point scale). ** and * denote statistically significant differences at the 1% and 5% levels, respectively, between first-time-meeting dialogues and family-included dialogues. †† indicates a statistically significant difference at the 1% level between individual interlocutors and family interlocutors in family-included dialogues (both assessed using the Mann-Whitney U test). Bold values indicate the highest value in each row. The average for family-included dialogues represents the average of individual interlocutors and family interlocutors.

4.1 Results of Post-Dialogue Questionnaire

Table 2 lists the results of the post-dialogue questionnaire. Although all items received relatively high ratings, averaging above 4 on a 5-point scale, first-time-meeting dialogues tended to receive higher ratings than family-included dialogues: a Mann-Whitney U test revealed significant differences between these dialogue types, with informativeness, comprehension, and familiarity exhibiting significance at the 1% level, and interest at the 5% level. A comparison between the individual interlocutors and family interlocutors in family-included dialogues indicated that individ-

ual interlocutors provided lower ratings, whereas family interlocutors' ratings were comparable to those observed in first-time-meeting dialogues. A Mann-Whitney U test conducted on these two groups showed significant differences at the 1% level across all items. This suggests that individual interlocutors may have experienced a sense of exclusion due to the other two interlocutors being family members, leading to lower evaluations.

On the basis of these results, we can infer that dialogues involving interlocutors with the same relationship type, such as first-time interlocutors in first-time-meeting dialogues or family interlocutors in family-included dialogues, tend to be rated more favorably by the interlocutors partaking in such a relation. This tendency may be explained by the similarity-attraction theory (Byrne and Nelson, 1965) or in-group bias (Brewer, 1979).

4.2 Analysis of Mention Usage

In multi-party dialogues, directing an utterance to a specific interlocutor is crucial for facilitating smooth conversation flow and turn-taking. In the text-based chat dialogues collected for this study, the mention feature explicitly indicates the addressee. The usage patterns of the mention feature—specifically, who is mentioned and how frequently—may differ between first-time-meeting dialogues and family-included dialogues due to varying relationships among participants, influencing the overall dynamics. This section quantitatively analyzes the differences in mention feature usage across dialogue types to examine the impact of relationship variations on addressee selection and communication styles.

We first compared the number of mention usages in first-time-meeting dialogues and family-included dialogues. The total number of mentions was 7,589 for first-time-meeting dialogues and 13,750 for family-included dialogues. A Mann-Whitney U test revealed that the number of mentions per dialogue was significantly higher in family-included dialogues than in first-time-meeting dialogues ($p < 0.01$). This suggests that communication explicitly addressing specific interlocutors was more active in family-included dialogues, which involve a mix of diverse relationships.

We further analyzed the directionality of mentions in family-included dialogues. The distribution of mentions was as follows: mentions between family interlocutors occurred 2,740 times (0.084 times per utterance), mentions from family interlocutors to an individual interlocutor occurred 5,185 times (0.16 times per utterance), and mentions from an individual interlocutor to family interlocutors occurred 5,796 times (0.29 times per utterance).[2] We conducted a Mann-Whitney U test to compare the number of mentions per dialogue for these directions and found that the number of mentions from family interlocutors to an individual interlocutor was significantly higher than the number of mentions between family interlocutors ($p < 0.01$). This suggests that family interlocutors actively used mentions to communicate with the individual interlocutor, rather than exclusively conversing among themselves. Furthermore, mentions from an individual interlocutor to family interlocutors were the most frequent across all directions. This indicates that the individual interlocutor was more engaged, either by actively seeking dialogue or by having greater opportunities to request information and express opinions to the family interlocutors.

These results suggest that, particularly in family-included dialogues where diverse relationships are intertwined, the mention feature not only served as a tool to indicate the addressee but also actively directed and facilitated interaction among multi-relational participants.

4.3 Specific Expressions for Each Type of Dialogue

In this section, we present the results of our investigation into linguistic expressions specific to first-time-meeting dialogues and family-included dialogues. The analysis was conducted from two perspectives: frequently occurring expressions and the linguistic cues used by humans to infer the relationships of interlocutors.

4.3.1 Analysis Based on Frequently Occurring Expressions

We extracted 4-gram patterns from utterances using the morphological analyzer MeCab[3](Kudo et al., 2004) with the IPAdic dictionary. Among the top 100 most frequently occurring 4-grams in the entire dialogue dataset, we identified and compared those with significantly different occurrence frequencies between first-time-meeting dialogues and family-included dialogues. Additionally, for

[2] In addition to mentions directed at a single interlocutor, there were 29 instances of mentions directed at two interlocutors, which were excluded from this analysis.

[3] https://taku910.github.io/mecab/

family-included dialogues, we identified and compared 4-grams with significantly different occurrence frequencies between utterances by the individual interlocutors and those by family interlocutors. The significance of these differences was assessed using a Z-test.

Our findings showed that, in first-time-meeting dialogues, greeting expressions such as "*Nice to meet you*" and agreement or empathy expressions like "*That's right*" and "*I see*" were frequently observed. In family-included dialogues, informal expressions such as "*Yeah, yeah*", "*Right!!!*", and "*...!*" were more prevalent. These expressions suggest a closer relationship among interlocutors. Focusing on the utterances of the individual interlocutors, question-related expressions such as "*Is it... ing?*" and "*Is there...?*" were frequently observed. This indicates a stronger tendency for individual interlocutors to seek information from family interlocutors, corroborating the findings from the mention analysis in Section 4.2, which revealed that mentions from individual interlocutors to family interlocutors were most frequent. In the utterances of family interlocutors, expressions in which handles were adapted into affectionate forms, such as "*[Adapted Handle] is...*" were observed, reflecting the closeness between family members. For actual dialogue examples, refer to Tables 3 and 4.

4.3.2 Analysis of Linguistic Cues Used by Humans to Infer Relationships of Interlocutors

To clarify which linguistic expressions suggest relationships among interlocutors, we asked crowdworkers[4] to select expressions they perceived as indicating relationships and to provide explanations for their corresponding reasons.

Workers were provided with information on the relationships among dialogue participants along with excerpts from the dialogue history. Each dialogue was divided into three segments: the beginning phase (utterances 1–33), the middle phase (utterances 34–66), and the ending phase (utterances 67–99). For each dialogue excerpt, three expressions and their corresponding reasons were collected. This task was conducted on ten first-time-meeting dialogues and ten family-included dialogues. A total of ten crowdworkers participated in the task, resulting in the collection of 90 expressions and reasons for each type of dialogue.

[4] https://crowdworks.jp/

The results showed that, in first-time-meeting dialogues, the lack of knowledge about basic information (e.g., one's occupation or place of residence), the use of a polite tone, and the shallow exploration of topics that fail to fully develop or become engaging were all cited as evidence indicating that these were first-time interlocutors. In contrast, in family-included dialogues, there were responses highlighting the difference in tone, such as the use of polite language with the individual interlocutor compared to more casual or familiar speech with the other family interlocutor. In addition, among the responses focusing on the family interlocutor, some cited evidence such as an exchange suggesting a long parent-and-child history, for example, "*(When you were a child) I wanted you to play hockey,*" or a moment where both family members reacted with exactly the same exclamation, "*That's amazing!*", suggesting an apparent familial bond. These various family-specific interactions were identified as cues that humans rely on when inferring relationships among interlocutors.

5 Evaluation of the Predictability of Relationships among Interlocutors

For a dialogue system to participate naturally in multi-party dialogues, understanding the relationships among interlocutors is crucial. Therefore, this section first investigates the extent to which humans can understand the relationships from the collected dialogues. For this purpose, we designed a relationship prediction task that assesses how accurately humans can infer the relationships among interlocutors when reading portions of dialogues. Specifically, we conducted experiments with two conditions: one restricting relationship options to two choices (first-time, family), and another expanded condition considering more fine-grained relationship diversity with three choices (first-time, acquaintance, family). Then, since clarifying how well LLMs can comprehend human relationships is essential to evaluate their current multi-party dialogue capabilities, we conducted the same task using LLMs.

5.1 Experimental Design

For evaluation, we randomly selected 50 first-time-meeting dialogues and 50 family-included dialogues. Following the task described in Section 4.3.2, each dialogue was divided into three segments: the beginning, middle, and ending.

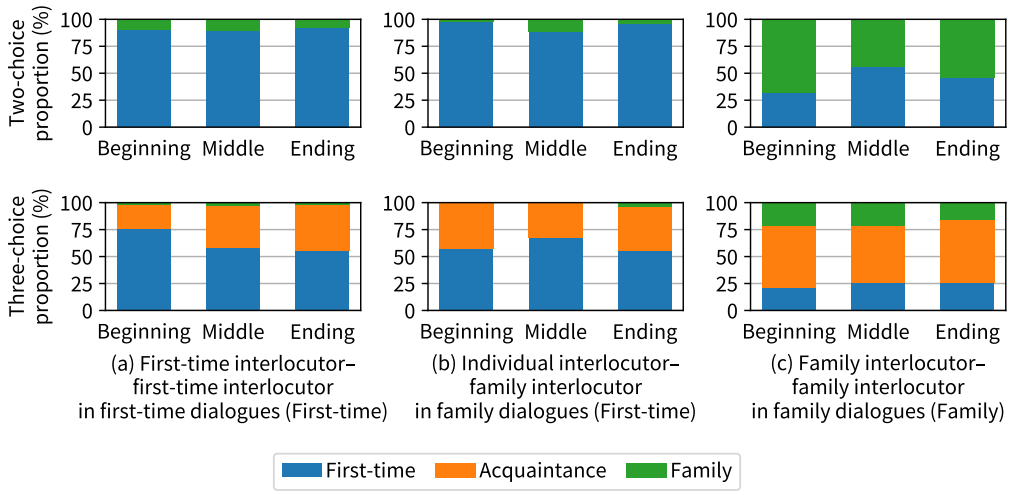

Figure 1: Prediction results for relationships of interlocutors by humans. The top and bottom rows show the results of a two-choice and three-choice classification, respectively. From left to right, the results correspond to the predicted relationships between (a) first-time interlocutors in first-time-meeting dialogues, (b) individual interlocutor and family interlocutor in family-included dialogues, and (c) family interlocutors in family-included dialogues.

Human evaluation was conducted by 51 crowdworkers for each of the 2-choice and 3-choice settings. Each worker was assigned three first-time-meeting dialogues and three family-included dialogues. For all six assigned dialogues, each worker performed tasks for one of three phases: beginning, middle, or ending. They predicted the relationships between all three pairs of participants (totaling three pairs) within each dialogue and provided reasons for their judgments. To prevent bias in the workers' judgments, we ensured that the same interlocutor did not appear across different dialogues within any worker's assigned set of six dialogues.

For the evaluation conducted by LLMs, we used the `gpt-4o-2024-08-06` model. The input provided to the LLM included the dialogue text and the names of the two interlocutors whose relationship was to be predicted, mirroring the input given to human evaluators. The model was then instructed to determine the relationship from the provided choices.

5.2 Human Prediction of Interlocutor Relationships

We present the results of our investigation into the extent to which humans can predict relationships between interlocutors from dialogues.

5.2.1 Prediction Results for 2-Choice Setting

The top row of Fig. 1 shows the results of the relationship prediction. In Fig. 1(a), which depicts dialogues between first-time interlocutors, the relationship between interlocutor pairs was correctly predicted as first-time with a high overall accuracy of approximately 90%. No significant decrease in accuracy was observed as the dialogue progressed, and consistently high accuracy was maintained from the beginning to the ending phase. Similarly, as shown in Fig. 1(b), relationships between individual and family interlocutors in family-included dialogues was predicted as first-time with a high overall accuracy exceeding 90%.

For relationships between family interlocutors, as shown in Fig. 1(c), the prediction accuracy for family remained around 55% overall. When examining the reasons why workers incorrectly predicted family relationships as first-time, some noted that their judgment was based on the absence of direct conversation between the two family interlocutors. Considering the analysis in Section 4.2, which revealed a higher frequency of mentions between first-time interlocutors and family interlocutors in family-included dialogues, it is possible that the limited number of interactions between family speakers, which could have provided clues for inferring their relationship, led to an increase in misclassifications by the workers. Furthermore, focusing on changes in prediction accuracy as the dialogue progressed, the accuracy in the beginning phase was higher compared to the middle and ending phases. This is likely because the gap in intimacy between the individual interlocutor and the family pair was more pronounced in the beginning

A:	Hello.
B:	Hello!
C:	Good morning.
B:	It's cold, isn't it?
C:	Yes, it is.
A:	It's still quite chilly.
B:	How's the weather?
C:	It's sunny.
A:	It's nice weather.
B:	Not a cloud in the sky.
C:	Same here.
B:	Were you busy this morning?
C:	I was nervous, so I woke up early.
A:	I'm not a morning person.

Table 3: Excerpt of a first-time-meeting dialogue. A, B, and C denote interlocutors.

D:	@F Were there any school subjects you liked?
F:	@D English and Japanese.
E:	Ah, so you're into humanities!
F:	But I can't speak English.
D:	@F I was bad at English$\cdots$
E:	I can't speak it either.
E:	Teba-san, you can speak a little, can't you?
D:	A little.
E:	Because you're a university student now!
E:	You have to do English compositions for your assignments, right?
F:	That sounds like you'd be able to speak it then.
D:	I did.

Table 4: Excerpt of a family-included dialogue. D, E, and F denote interlocutors. Interlocutors D and E are family (E is the mother and D is her daughter). Note that *Teba-san* is the nickname derived from D's handle name.

phase of the dialogue.

These results from the 2-choice experiments indicate that interlocutor relationships can be identified to some extent from dialogue content. This strongly suggests that the dialogues in the MRMP corpus constructed in this study indeed contain features that express the relationships between interlocutors. While identifying family relationships was not as straightforward as identifying first-time interlocutor relationships, our findings suggest that clues are expressed, particularly in the beginning phase of the conversation.

5.2.2 Prediction Results for 3-Choice Setting

Considering that actual human relationships are more diverse than a simple binary classification of first-time or family, we conducted an experiment to predict relationships using a 3-choice setting (first-time, acquaintance, family), which includes acquaintance as an intermediate category. The experimental procedure, dialogue data for evaluation, and dialogue segmentation method were identical to those used in the 2-choice setting. The evaluation was conducted by 51 crowdworkers newly recruited through crowdsourcing.

The bottom row of Fig. 1 shows the results of the relationship prediction. In Fig. 1(a), the proportion of correctly predicted first-time relationships between first-time interlocutors decreased to approximately 60% overall, with most misclassifications being categorized as acquaintance. Focusing on the dialogue's progression, while the prediction accuracy for first-time was highest in the beginning phase, the proportion judged as acquaintance tended to increase as the dialogue advanced into the middle and ending phases.

In family-included dialogues, the prediction accuracy for the first-time relationship between the individual interlocutor and family interlocutors (Fig. 1(b)) also decreased compared to the 2-choice setting. While there was a tendency for first-time predictions to increase during the middle phase of the dialogue, instances of being judged as family also emerged in the ending phase.

For the prediction of relationships between family interlocutors (Fig. 1(c)), the overall accuracy for predicting family decreased significantly to approximately 20% from about 55% in the 2-choice setting. In the 3-choice setting, the notable increase in prediction accuracy for family relationships observed in the beginning phase during the 2-choice setting was not present. Instead, accuracy remained consistently low regardless of the dialogue's progression phase. Most misclassifications involved predicting family as acquaintance. Comparing the 2-choice and 3-choice prediction results, it seems that when a certain degree of relationship was recognized, but without enough certainty to be identified as family, it was categorized as acquaintance. This might explain why some instances previously predicted as family in the 2-choice setting were predicted as acquaintance in the 3-choice setting.

5.3 LLM Prediction of Interlocutor Relationships

We tested GPT-4o, an LLM, to clarify its prediction tendencies on tasks performed by humans.

5.3.1 LLM Prediction Results for 2-Choice Setting

We conducted a prediction experiment using GPT-4o in a 2-choice setting, where the model selected the relationship between interlocutors as either first-time or family. The results are shown in the top

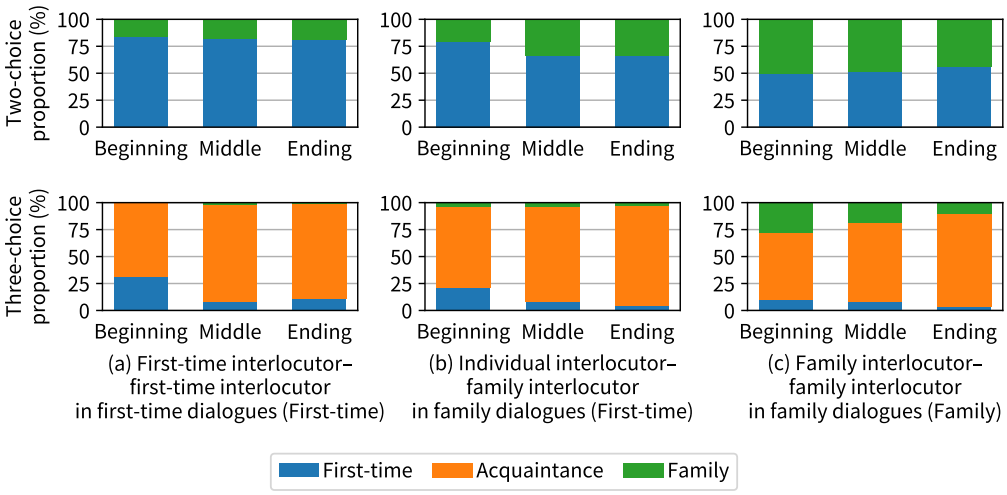

Figure 2: Prediction results for relationships of interlocutors by GPT-4o. The top and bottom rows show the results of a two-choice and three-choice classification, respectively. From left to right, the results correspond to the predicted relationships between (a) first-time interlocutors in first-time-meeting dialogues, (b) individual interlocutor and family interlocutor in family-included dialogues, and (c) family interlocutors in family-included dialogues.

row of Fig. 2. In first-time-meeting dialogues, prediction accuracy for first-time relationships was approximately 80%. This was slightly lower than the accuracy achieved by human evaluators under the same conditions (approximately 90% overall).

Regarding the relationship between the individual interlocutor and family interlocutors in family-included dialogues, the accuracy was around 70%, which also fell short of human performance.

For relationships between family interlocutors in family-included dialogues, the accuracy was approximately 50%. Overall, while GPT-4o did not match human accuracy, it demonstrated a certain level of predictive ability.

5.3.2 LLM Prediction Results for 3-Choice Setting

We conducted a prediction experiment using GPT-4o with acquaintance added to the relationship choices. The results are shown in the bottom row of Fig. 2. In the 3-choice setting, GPT-4o's prediction tendencies changed significantly from the 2-choice setting. For first-time interlocutor relationships (see Fig. 2(a), (b)), GPT-4o categorized approximately 80% as acquaintance throughout, while the proportion correctly predicted as first-time remained relatively low.

For predicting relationships between family interlocutors, GPT-4o's highest proportion of judgments, approximately 75%, was acquaintance, while the proportion correctly predicted as family was about 20%. Although human accuracy was also low under these conditions (around 20%), GPT-4o exhibited a stronger bias towards acquaintance than human evaluators.

Summarizing these results, in the 3-choice setting, GPT-4o exhibited significantly different prediction tendencies compared to human evaluators. Across most dialogue types and phases, a strong bias was observed towards classifying relationships as acquaintance, which resulted in decreased accuracy for correctly identifying first-time or family relationships. This tendency suggests that GPT-4o may have insufficient recognition of expressions specific to first-time and family relationships, indicating that it cannot clearly distinguish these from the more general category of acquaintance.

6 Conclusion

In this study, with the aim of developing a multi-party dialogue system, we constructed the Multi-Relational Multi-Party Chat Corpus (MRMP), a multi-party dialogue corpus that focuses on the relationships of interlocutors. Using this corpus, we conducted analyses to elucidate the impact of the relationships of interlocutors on multi-party dialogue. Additionally, through the implementation of relationship prediction tasks, we identified challenges for LLMs in recognizing interlocutor relationships.

As future work, we plan to utilize this corpus and the insights obtained here to develop a multi-party dialogue system. In addition, since the relationships covered by this corpus are limited to first-

time interlocutors and family relationships, we also intend to collect dialogues among acquaintances to gain further insights. Furthermore, since relationships are often conveyed through speech and other modalities, we plan to collect and analyze data from spoken and multimodal multi-party interactions as well.

Acknowledgments

This work was supported by JST Moonshot R&D Grant number JPMJMS2011.

References

Marilynn B Brewer. 1979. In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological bulletin*, 86(2):307–324.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Donn Byrne and Don Nelson. 1965. Attraction as a linear function of proportion of positive reinforcements. *Journal of personality and social psychology*, 1(6):659–663.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Proceedings of the International workshop on machine learning for multimodal interaction*, pages 28–39.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 610–614.

Yasuharu Den and Mika Enomoto. 2007. A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. *Conversational informatics: An engineering approach*, pages 305–330.

Yasuharu Den, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida. 2010. Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2103–2110.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1597–1601.

Shinya Iizuka, Shota Mochizuki, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, and Ryuichiro Higashinaka. 2023. Clarifying the dialogue-level performance of GPT-3.5 and GPT-4 in task-oriented and non-task-oriented dialogue systems. In *Proceedings of the AAAI Symposium Series*, volume 2, pages 182–186.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal*, volume 1, pages 364–367.

Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. 2022. Design and evaluation of the corpus of everyday Japanese conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5587–5594.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Khyati Mahajan and Samira Shaikh. 2021. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: A deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Katsuya Takanashi and Hitoshi Isahara. 2002. San-sha kaiwa deta no shuroku hoho oyobi bunseki wakugumi no gaiyo. In *Proceedings of the 8th annual meeting of the Association for Natural Language Processing*, pages 116–119.

Kazuya Tsubokura, Rina Kato, and Kunikazu Kobayashi. 2022. Field test on multi-party dialogue robot—analysis of parent-child conversation in dialogue with robot—. In *JSAI Technical Report, SIG-Challenge-060*, pages 1–6. (In Japanese).

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934.