Estimating Relationships between Participants in Multi-Party Chat Corpus

Anonymous ACL submission

Abstract

While most existing dialogue studies focus on dyadic (one-on-one) interactions, research on multi-party dialogues has gained increasing importance. One key challenge in multi-party dialogues is identifying and interpreting the relationships between participants. This study focuses on multi-party chat corpus and aims to estimate participant pairs with specific relationships, such as family and friends. The proposed model extracts features from the input text, including the number of turns and the frequency of honorific expressions, and trains a logistic regression model to predict relationships. Experiments demonstrated that the proposed model significantly outperforms LLM in relationship estimation tasks.

1 Introduction

In multi-party dialogues, we naturally infer human relationships and degrees of intimacy among participants, and adapt our linguistic choices and social behaviors accordingly. The recognition of these interpersonal relationships plays a crucial role in facilitating smooth and effective communication.

Compared to dyadic dialogue, the structure and characteristics of multi-party dialogue have been less extensively studied. This is primarily due to the challenges associated with recording and analyzing multi-party conversations. For example, Ishizaki and Kato (1998) examined the characteristics of three-party dialogues in comparison to two-party dialogues, and found that even in three-party settings, a significant amount of information tends to be exchanged predominantly between two participants. Additionally, according to Clark (1982), the role of listeners is not uniform in dialogues involving three or more participants. He argued that, in response to a speaker's utterance, some participants take the role of addressees while others become collateral participants, and that speakers employ various means to designate these roles during the conversation. Novick et al. (1970) also discusses the unique characteristics of multi-party dialogues. Specifically, they pointed out that: (1) sub-dialogues may emerge among a subset of participants to serve the interests or purposes of a specific group; (2) multiple listeners may collaborate in supporting or responding to the speaker's communicative efforts; and (3) in multi-party dialogues, new utterances may overwrite previous ones, leaving some utterances unacknowledged. As these studies suggest, the analysis of multi-party dialogue is inherently more complex than that of two-party dialogue. In recent years, large language models have also been applied to multi-party dialogue tasks. Tan et al. (2023), for instance, employed GPT-4 for five representative tasks: emotion detection, speaker identification, destination identification, response selection, and response generation, but their performance is limited.

Moreover, many studies on relationship recognition have focused primarily on bilateral interactions. For example, several systems have been proposed to estimate the degree of rapport or hierarchical relationships based on the types of attitudes a speaker adopts toward a listener (Nishihara et al., 2008), or to predict the level of intimacy based on the number of sentences and the frequency of positive emotional expressions (Matsumoto et al., 2018). However, since these methods are designed under the assumption of two-party dialogues or rely on scripted dialogues, their applicability to natural multi-party conversations remains unclear.

In this study, we propose a method for estimating interpersonal relationships among participants using multi-party dialogues. The objective is to identify both the types and the depth of the relationships. According to Matsumoto et al. (2005), interpersonal relationships are reflected in the content of dialogue. When humans infer relationships from conversations, they utilize various cues such

as speech content, eye contact, smiles, and gestures. Among these, it has been reported that approximately 40% of the information needed to estimate interpersonal relationships is derived from the speech content alone (Kimura, 2006). Based on these, it is presumed that the content of textual dialogue is the most important factor for dialogue systems to infer the relationships among participants. The proposed method estimates interpersonal relationships among participants by extracting features such as the number of turns, the frequency of honorific expressions, the frequency of questions, and the use of mention tags (i.e., the explicit designation of the recipient of a turn). We also compared the performance of the proposed method with that of GPT-4, a widely used large-scale language model, using zero-shot and few-shot prompting for the relationship identification task. In addition to standard prompts, we also experimented with prompts that asked GPT to output the reasoning behind its estimations in some tasks, allowing us to examine how GPT justifies its relationship inferences.

2 Multi-Party Chat Corpus

This section provides an overview and examples of the corpus used in this study and the preprocessing that was applied.

2.1 Corpus Overview

The multi-party chat corpus used in this study was developed by Tsuda et al. (2025) and consists of text-based three-party dialogues in Japanese. The participants chatted in text online for approximately 100 turns after they were invited to the online meeting space. They chatted in Japanese. Each dialogue was terminated at a natural topic boundary after it exceeded 100 turns. Here, a unit that ends with a line break is counted as a separate turn.

The dialogues are broadly categorized into three types based on the relationships among the three participants: dialogues among strangers (meeting for the first time), dialogues involving two family members and one stranger, and dialogues involving two acquaintances and one stranger. The first type will be referred to as "first-time dialogue," the second as "family dialogue," and the third as "acquaintance dialogue." The participants consist of six family pairs (12 participants), a group of 16 mutual acquaintances, and 115 participants who were complete strangers. Each turn is annotated with the speaker, the utterance content, and, when the speaker wants to, a mention tag (@name) explicitly indicating the intended addressee. Each group of participants engaged in five or ten dialogues; each dialogue was conducted independently, and the discussion topics were not shared across dialogues.

Speaker	Utterance
A	Did you have breakfast this morning?
B	@A Yes, I did!
C	@A I had soba!
B	@A The green onions were spicy in mom's natto rolls.
A	@C Looks great for the morning!
C	@B Natto rolls!
A	I had to make 6 natto rolls. For three people.
C	That sounds like a lot of effort!

Table 1: Example of chat corpus (family dialogues; translated from original Japanese. "@" represents a mention tag.)

The corpus contains 1,000 first-time meeting dialogues, 500 family dialogues, and 500 acquaintance dialogues. An example of a family dialogue is presented in Table 1. From this example, we can easily infer the relationships among the participants: Speaker A is Speaker B's mother.

2.2 Preprocessing

For the corpus used in this study, we prepared three types of datasets, as shown below, by applying processing related to mention tags. Since the criteria for assigning mention tags can vary across participants, relying on human annotation alone may lead to inconsistencies. To address this issue, we prepared two versions of the corpus: one with all mentions removed and another with mentions automatically estimated.

- Original data
- Data without mention tags (by removing them)
- Data with estimated mention tags (by predicting them)

First, we conducted experiments using the original corpus data, as shown in Table 1

Second, we created a version of the corpus with the mention tags removed.

Input	
Speaker	Utterance
A	I've been immersed in baseball with my kids.
B	That's nice!
C	Sounds great!
C	You even play catch when you go home during the week, right?

Task	Correct Output Example
R	acquaintance dialogue
RP	A and C
R and P	acquaintance: A and C
RD	1

Table 2: Input and Output Example (acquaintance dialogue, R: Relationship, RP: Relational Pair, R and P: Relationship and Pair, RD: Relationship Depth)

Third, we created a version with automatic mention tags assignment for all turns using GPT. Specifically, we provided GPT-4o with a sequence of 10 turns, and for the final turn, we asked it to estimate the mention tag as either "@A", "@B", "@C", or "@all". To obtain stable outputs, a few-shot prompt was used. This process was applied to all turns, resulting in the creation of a chat corpus with mention tags for all turns. A number of studies have been conducted on the addressee recognition (AR) task (e.g. Le et al., 2019; Li and Zhao, 2023; Tan et al., 2023), and according to Tan et al. (2023), the correct response rate for GPT-4 in the AR task is 82.5%. For the corpus used in this study, the correct response rate was 65.2%. A large difference in performance is that the dataset used by Tan et al. was from the Ubuntu IRC, which mainly consists of questions and answers, and is different from the casual conversation used in this study.

3 Task Definition

In this study, the following tasks were defined to examine the relationships among the participants in the multi-party dialogue. Examples of the input and output for each task are also shown in Table 2. We provided the whole dialogue (approximately 100 turns) in a prompt. However, to illustrate how different outputs are generated depending on the task, we present only four turns here as an example.

3.1 Relationship Identification Task (R)

The relationship identification task is defined as a three-class classification task aimed at determining the dialogue type based on participant relationships, as mentioned in Section 2.1: First-time, Family, and Acquaintance dialogues.

3.2 Relational Pair Identification Task (RP)

The relational pair identification task focuses on Family and Acquaintance dialogue. This classification task aims to identify the pair of speakers of family members or acquaintances, respectively. Here, the task is performed for given dialogues consisting of two family-or-acquaintance participants and one stranger. The characteristics of pairwise estimation for each relationship are analyzed.

3.3 Relationship and Pair Identification Task (R and P)

This task is a combination of the two tasks mentioned above, that is to identify the two participants with a relationship in family and acquaintance dialogues, and simultaneously determine whether they are a family pair or an acquaintance pair. The simultaneous estimation of both the relationship and the pair will facilitate its application to dialogue systems.

3.4 Relationship Depth Assessment Task (RD)

Each group of participants was engaged in five or more dialogue sessions. We also expect the difference between the first and fifth sessions in terms of the depth of the relationship. Hayashi et al. (2023) define rapport as the feeling of connection and harmony with the other person, showing that rapport increases as the number of conversations grows. Therefore, a higher rapport or a deeper relationship, the depth of the relationship is expected to emerge in the fifth session compared to the first session. In the relationship depth assessment task, we focus on data from the first and fifth dialogues with the same participants, and identify whether the dialogue is the first or fifth one.

4 Analysis and Proposed Method

4.1 Statistical Analysis

We investigated features (referred to as "dialogue features"), including the number of turns, honorifics, questions, and mention tags, derived from sentences proposed by Matsumoto et al. (2018).

Among these, the number of honorifics was measured using a rule-based approach, with expressions such as "desu" and "masu" [1] being each counted as one instance. Similarly, the number of questions was also measured using a rule-based method, counting each question as a single occurrence. Three types of mention-related features were measured:

- The number of mention tags used from each participant to each other participant
- The number of mention tags with honorifics from each participant to each other participant
- The number of mention tags with questions from each participant to each other participant

They were adopted based on the assumption that identifying relationships would be easier by focusing on the addressee of honorifics and questions. All features were measured by absolute counts per dialogue, but all dialogue sessions consist of approximately 100 turns.

The results of the analysis for each type of dialogue and participant type are shown in Table 3. Table 4 also presents the results for mention-related features. Table 5 shows the results of *t*-tests comparing the mean differences between first-timers and family members, and between first-timers and acquaintances. The results reveal that there are statistically significant differences at the 0.05 significance level between the first-time participant and family members, and also between the first-time participant and acquaintances. They were observed between the following pairs: (1) family member (mean per person) and first-timer in family dialogues, and (2) acquaintance (mean per person) and first-timer in acquaintance dialogues. This suggests that in three-party dialogues, when a family pair is present, fewer direct conversations occur between them, while when an acquaintance pair is present, more conversations take place between them. This relationship is illustrated in Figure 1. This may be attributed to the fact that, in the family dialogues, the conversation tended to evolve around the first-timer, whereas in the acquaintance dialogues, the two acquaintances often became more engaged with each other and carried on the conversation more actively between themselves.

[1] In Japanese, "desu" and "masu" are commonly used to express politeness and respect, which are part of the honorifics system in the language.

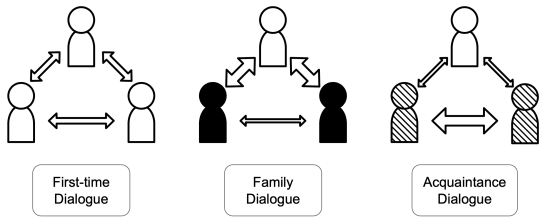

Figure 1: Interaction patterns in first-timer, family, and acquaintance dialogues. The white person represents the first-timer, the black person represents a family member, and the striped person represents an acquaintance. Arrow thickness indicates the frequency of interaction between each pair.

Dialogue Type	Partici-pants	#Uttera-nces	#Honori-fics	#Ques-tions
First-time	First	34.4 (8.0)	20.6 (7.9)	3.2 (2.7)
Family	First	39.8 (7.2)	22.9 (7.4)	7.2 (4.5)
Family	Family	32.6 (7.8)	16.5 (6.6)	3.0 (2.4)
Acq	First	27.9 (7.9)	13.1 (7.3)	4.0 (3.0)
Acq	Acq	39.3 (10.8)	8.5 (5.6)	4.4 (3.2)

Table 3: Mean values (and standard deviations) per participant for each dialogue type and feature (First: first-timer, Acq:acquaintance)

4.2 Proposed Method

Based on the results of these analyses, the proposed method performs logistic regression using the dialogue features extracted from the sentences. To evaluate the proposed method, we employed 10-fold cross-validation.

- Number of turns per participant
- Number of honorifics per participant
- Number of mention tags per participant
- Number of mention tags with honorifics per participant
- Number of mention tags with questions per participant

However, since the number of mention tags could not be measured in the dataset where mention tags were removed, we did not use any mention-related features.

5 Evaluations

Each of the tasks described in Section 3 was evaluated using the methods proposed in Section 4. For reference, we employ GPT-4o using both zero-shot and few-shot approaches. In addition, we also experiment with GPT-4o prompts that incorporate

Speaker	Mentioned person	#Mention tags	#Mention tags /w honorifics	#Mention tags /w questions
First-timer	First-timer	4.0	2.9	0.5
First-timer	Family	5.8	3.7	1.3
Family	First-timer	5.2	4.1	0.9
Family	Family	2.7	0.3	0.4
First-timer	Acquaintance	5.4	2.9	0.9
Acquaintance	First-timer	8.0	4.0	1.2
Acquaintance	Acquaintance	9.1	0.4	1.2

Table 4: Mean values of mention-related dialogue features

Participants	Dialogue Features	t-value
First-timer and Family (@family dialogue)	#Turns	17.8*
	#Questions	19.0*
	#Honorifics	16.4*
	#Mention tags	17.4*
	#Mention tags /w honorifics	31.0*
	#Mention tags /w questions	14.7*
First-timer and Acquaintance (@acq dialogue)	#Turns	**23.1***
	#Questions	**2.4***
	#Honorifics	12.4*
	#Mention tags	**12.8***
	#Mention tags /w honorifics	27.9*
	#Mention tags /w questions	**3.1***

Table 5: t-test results (two-tailed), The asterisk (*) denotes significance with $p < 0.05$. The t-value is bolded when Family > First-timer or Acquaintance > First-timer. (acq: acquaintance)

statistical properties of conversational structure, which were identified through Section 4.1.

5.1 Relationship Identification Task

The following is the prompt for the GPT-based method for the relationship identification task.

> Analyze a conversation between three people and output in one line an estimate of whether it includes two family members, or two people who are not family members but who know each other, or whether no one is family or knows each other.
> The output format should be "family" only if family pairs are included. If acquaintance pairs are presumed to be included, output only "acquaintances". If neither family nor acquaintances are presumed, output only "no".

For the few-shot prompting, we provided several appropriate examples depending on the type of dialogue and whether mention tags were present. For the few-shot and statistical information prompt, we incorporated several statistical properties observed in our analysis, such as: honorifics are rarely used among family members or acquaintances; utterances between family members are infrequent; and utterances between acquaintances are more frequent. In order to estimate the relationship, the proposed method employs logistic regression with three class categories: first-time dialogue, family dialogue, and acquaintance dialogue.

The results are presented in Table 6. The evaluation metrics shown in Table 6 represent macro-averages. The results indicate that the proposed method achieved the highest performance on the original data. Additionally, all of the proposed methods outperformed GPT. In the zero-shot prompts of GPT, there was a large difference between precision and recall depending on the class, resulting in an extremely low F1-score. For both the proposed method and GPT, the performance did not significantly decline when using the dialogues without mention tags. However, the performance dropped when using the dialogues with estimated mention tags. This is likely because the mention tags prediction accuracy was low, but they were added to all turns. In the prompts augmented with statistical information, a slight improvement was observed only when using the dialogues without mention tags or with estimated mention tags.

5.2 Relational Pair Identification Task

The following is the prompt for the GPT-based method for the relational pair identification task. The term "family" was replaced with "acquaintance" in the experiments involving acquaintance dialogues.

Method	Precision	Recall	F1-score
GPT-4o w/ M	0.54	0.51	0.41
GPT-4o w/o M	0.48	0.47	0.42
GPT-4o w/ EM	0.48	0.42	0.31
GPT-4o FS w/ M	0.62	0.61	0.61
GPT-4o FS w/o M	0.47	0.47	0.47
GPT-4o FS w/ EM	0.50	0.49	0.49
GPT-4o FS+ST w/ M	0.61	0.61	0.61
GPT-4o FS+ST w/o M	0.58	0.51	0.51
GPT-4o FS+ST w/ EM	0.55	0.53	0.54
Proposed w/ M	**0.80**	**0.78**	**0.79**
Proposed w/o M	0.79	0.77	0.78
Proposed w/ EM	0.75	0.73	0.73

Table 6: Results of Relationship Identification Task (FS: Few-Shot, ST: Statistic, M: Mention, EM: Estimated Mention)

Method	Family	Acquaintance
GPT-4o w/ M	0.44	0.72
GPT-4o w/o M	0.44	0.70
GPT-4o w/ EM	0.35	0.61
GPT-4o FS w/ M	0.64	0.70
GPT-4o FS w/o M	0.66	0.69
GPT-4o FS w/ EM	0.51	0.67
GPT-4o FS+ST w/ M	0.69	0.68
GPT-4o FS+ST w/o M	0.59	0.65
GPT-4o FS+ST w/ EM	0.59	0.69
Proposed w/ M	**0.99**	**0.98**
Proposed w/o M	0.88	0.92
Proposed w/ EM	**0.99**	0.97

Table 7: Relational Pair Identification Task (Accuracy, FS: Few-Shot, ST: Statistic, M: Mention, EM: Estimated Mention)

> Analyze the conversation and estimate which two of the three are the family pair.
> The output format should be only "A and B", for example, if you think that A and B are a family pair.

Furthermore, for a subset of the first-time and family dialogue data, the prompts were designed to elicit outputs that explicitly include the reasoning process.

The proposed method employed logistic regression with three classification targets: A and B, A and C, and B and C. The results for the family and acquaintance dialogues are presented in Table 7, showing the percentage of correctly identified pairs. A summary of the output reasons, including the inference process, is provided in Table 8.

According to Table 7, for both family and acquaintance relationships, the proposed method using the original dataset achieved the highest accuracy. In the zero-shot prompting, GPT performed better on acquaintance dialogues than on family dialogues. However, in the few-shot prompting, the performance on family dialogues improved, reducing the gap between the two types of dialogue. In the prompts augmented with statistical information, no consistent improvement was observed, as the performance varied depending on the method. According to Table 8, in the zero-shot prompting, incorrect predictions were often made by empathy or frequent interactions. Table 9 presents an example where GPT made an error in pair estimation: in this case, although the correct answer was B and C, GPT incorrectly inferred that A and B formed the family pair, reasoning that they were empathizing with each other over a topic related to children. This suggests that LLMs tend to interpret close communication—such as frequent exchanges—as indicative of a close relationship. As discussed in the analysis in section 4, acquaintance dialogues contain more exchanges between the acquaintances themselves, which may explain why GPT produced better results for acquaintance dialogues than for family dialogues. However, with few-shot prompting, fewer incorrect predictions were attributed to factors such as calling by name, empathy, or frequent interactions. Unlike the Relationship Identification Task, the performance decreased when using the dialogues without mention tags, while it improved when using the dialogues with estimated mention tags. This suggests that mention-related features have a strong impact on identifying rela-

Reason	C	IC	C (FS)	IC (FS)
Calling by Name or Relationship	27	41	24	15
Empathy	1	36	5	20
Frequent Interactions and Questions	0	28	0	12
Shared Topics	82	6	106	15
Others	20	9	35	18

Table 8: Reasons for GPT's Family Pair Identification (C: Correct, IC: Incorrect, FS: Few-Shot)

Speaker	Utterance
A	That's why when I go to a big store, I end up taking my time looking around.
B	@A That's so true! When you have kids with you, you can't really take your time.
B	I quickly go while they're at school!
A	Yeah, definitely hard to take it slow with kids.
A	That's a good idea.
C	@B It's true, you can't really take your time.

Table 9: Example of data where GPT made an error (family dialogues)

Method	Accuracy
GPT-4o w/ M	0.34
GPT-4o w/o M	0.34
GPT-4o w/ EM	0.22
GPT-4o FS w/ M	0.40
GPT-4o FS w/o M	0.44
GPT-4o FS w/ EM	0.33
GPT-4o FS+ST w/ M	0.44
GPT-4o FS+ST w/o M	0.45
GPT-4o FS+ST w/ EM	0.40
Proposed w/ M	**0.95**
Proposed w/o M	0.78
Proposed w/ EM	0.90

Table 10: Relationship and Pair Identification Task (FS: Few-Shot, ST:Statistic, M: Mention, EM: Estimated Mention)

tionship pairs, and that predicted mention tags with low accuracy were effective to some extent.

5.3 Relationship and Pair Identification Task

The following is the prompt for the GPT-based method for the Relationship and Pair Identification task.

> Analyze a conversation between three people and estimate which two of the three are a related pair and what kind of relationship they have and output in one line.
> The output format should only be "Family: A and B" if family pairs are included. If the pair is not a family but an acquaintance, output only "Acquaintance: A and B".

In the proposed method, relationship–pair estimation was performed using two classes for relationship type (family or acquaintance) and three classes for pair combinations, resulting in a logistic regression model with six classification categories. The experimental results are presented in Table 10. The table shows the percentage of correct answers where both the relationship type and the specific pair were correctly identified.

It shows that the proposed method using the original data achieved the highest percentage of correct answers. The performance decreased when using the dialogues without mention tags, while it improved when using the dialogues with estimated mention tags. This is likely because, while the accuracy using the dialogues with estimated mention tags declined in the Relationship Identification task, the improvement in the accuracy in the Relational Pair Identification task was more substantial. In the prompts augmented with statistical information, overall performance improved.

5.4 Relationship Depth Assessment Task

The following is the prompt for the GPT-based method for the relationship depth assessment task.

> Analyze the conversation and output 1 or 5 for the dialogue, whether it is the first or fifth dialogue. The 1st and 5th dialogue data are given. The output format should be "numeric" only.

The proposed method performed a two-class logistic regression for estimating the depth of the relationship, classifying the first and fifth dialogue. The experimental results for first-time dialogues, family dialogues, and acquaintance dialogues are shown in Table 11.

According to Table 11, the proposed method achieved the highest accuracy in first-time dialogues, whereas GPT showed the highest accuracy in both family and acquaintance dialogues. In this task, the overall performance was low, even though it was a binary classification problem, and regardless of whether mention tags were present or not. The dialogue features used in the logistic regression model for estimating the depth of relationships were insufficient. Moreover, the use of mention tags tends to vary greatly depending on the individual, and it is likely that the mention-related fea-

Method	First-timer	Family	Acquaintance
GPT-4o w/ M	0.46	0.53	0.51
GPT-4o w/o M	0.51	0.54	0.54
GPT-4o w/ EM	0.50	0.50	0.58
GPT-4o FS w/ M	0.53	0.77	**0.70**
GPT-4o FS w/o M	0.53	**0.79**	**0.70**
GPT-4o FS w/ EM	0.52	0.76	0.65
Proposed w/ M	0.63	0.68	0.66
Proposed w/o M	0.64	0.71	0.65
Proposed w/ EM	**0.66**	0.70	0.62

Table 11: Relationship Depth Assessment Task (Accuracy, FS: Few-Shot, M: Mention)

tures did not change significantly between the first and fifth dialogues. The definition of relationship depth—based on whether the conversation was the first or fifth interaction—may have been inadequate. It is possible that even by the fifth conversation, the relationship had not deepened significantly enough to be effectively captured by the model.

6 Conclusions

In this study, we aimed to estimate inter-personal relationships in multi-party conversations by proposing a logistic regression model based on dialogue features. The effectiveness of the proposed method was evaluated through comparative experiments with GPT-4o. The analysis confirmed that first-time interactions, family conversations, and conversations between acquaintances each exhibit distinctive characteristics. By incorporating these findings into its feature design, the proposed method achieved significantly higher accuracy than GPT across multiple tasks, including detecting the presence of relationships and identifying specific relationship pairs. In particular, the proposed method showed outstanding performance in the relationship pair identification task. This is because the mention tags contributed significantly to relational pair identification. In contrast, in other tasks, the accuracy of mention tag prediction was lower, resulting in worse performance compared to the dialogues without mention tags. GPT tends to emphasize frequent and dense communication, resulting in relatively good performance for acquaintance conversations in the pair identification task, but showing lower accuracy for family conversations. These findings suggest that GPT is relatively capable of estimating the depth of relationships, despite its limitations in accurately identifying specific relationships. The findings of this study provide new metrics and methodologies for relationship estimation, contributing to the development of more adaptive and context-sensitive dialogue systems and the design of conversations tailored to interpersonal dynamics.

Future challenges include generalizing the model using diverse datasets, such as the Corpus of Everyday Japanese Conversation (CECJ, Koiso et al., 2022). Additionally, since the culture of honorifics is unique to Japan, it is essential to explore how this cultural aspect is represented in other languages. Furthermore, improvements to the model are needed to capture an even broader range of relationship depths and types. Moreover, if addressee information can be captured more accurately, the model's performance is expected to improve significantly. Accurate addressee recognition is crucial.

Limitations

This study was conducted using a Japanese dialogue dataset and is based on the linguistic and conversational characteristics specific to the Japanese language. For broader application and generalization, further validation is required using datasets from other languages and cultural contexts.

References

H. H. Clark. 1982. Hearers and speech acts. *Language*, pages 332–373.

Takato Hayashi, Ryusei Kimura, Ryo Ishii, Fumio Nihei, Atsushi Fukayama, and Shogo Okada. 2023. Ranking conversations based on rapport in first meeting conversations and friend conversations. In *SIGSLUD*, pages 72–79.

Masato Ishizaki and Tsuneaki Kato. 1998. Exploring the characteristics of multi-party dialogues. In *Association for Computational Linguistics*, page 583–589.

Masaki Kimura. 2006. Clarification of the cognitive mechanisms of interpersonal communication as a social skill. In *47th Annual meeting of the Japanese Society of Social Psychology*, pages 122–123.

Hanae Koiso, Haruka Amatani, Yuichi Ishimoto, Yuriko Iseki, Yasuyuki Usuda, Wakako Kashino, Yoshiko Kawabata, Yayoi Tanaka, Yasuharu Den, Kenya Nishikawa, and Yuka Watanabe. 2022. Design and features of the corpus of everyday japanese conversation. In *NLP*, page 2008–2012.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who

is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 1909–1919.

Yiyang Li and Hai Zhao. 2023. Em pre-training for multi-party dialogue response generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 92–103.

K. Matsumoto, J. Minato, F. Ren, and S. Kuroiwa. 2005. Estimating human emotions using wording and sentence patterns. In *International Conference on Information Acquisition (IEEE)*, pages 421–426.

Kazuyuki Matsumoto, Kyosuke Akita, Ren Fuji, Minoru Yoshida, and Kenji Kita. 2018. Intimacy estimation of the characters in drama scenario. *Intelligence and Information*, pages 591–604.

Yoko Nishihara, Wataru Sunayama, and Masahiko Yachida. 2008. Human friendship and hierarchical relationship estimation from utterance texts. *The Institute of Electronics, Information and Communication Engineers Transactions. Information and Systems: D*, pages 78–88.

David Novick, Lisa Walton, and Karen Ward. 1970. Contribution graphs in multiparty discourse. In *International Symposium on Spoken Dialogue (ISSD)*, pages 53–56.

Chao-Hong Tan, Jia-Chen Gu, and Zhen-Hua Ling. 2023. Is chatgpt a good multi-party conversation solver? In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Taro Tsuda, Sanae Yamashita, Koji Inoue, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2025. Multi-relational multi-party chat corpus. In *NLP*.