

Data Augmentation Integrating Dialogue Flow and Style to Adapt Spoken Dialogue Systems to Low-Resource User Groups

Zhiyang Qi

The University of
Electro-Communications
1-5-1, Chofugaoka, Chofu,
Tokyo, Japan
qizhiyang@uec.ac.jp

Michimasa Inaba


The University of
Electro-Communications
1-5-1, Chofugaoka, Chofu,
Tokyo, Japan
m-inaba@uec.ac.jp

Abstract

This study addresses the interaction challenges encountered by spoken dialogue systems (SDSs) when engaging with users who exhibit distinct conversational behaviors, particularly minors, in scenarios where data are scarce. We propose a novel data augmentation framework to enhance SDS performance for user groups with limited resources. Our approach leverages a large language model (LLM) to extract speaker styles and a pre-trained language model (PLM) to simulate dialogue act history. This method generates enriched and personalized dialogue data, facilitating improved interactions with unique user demographics. Extensive experiments validate the efficacy of our methodology, highlighting its potential to foster the development of more adaptive and inclusive dialogue systems.

1 Introduction

As an innovative technology at the forefront of artificial intelligence and speech processing, spoken dialogue systems (SDSs) have attracted significant interest from both academia and industry (Kawahara, 2018; Si et al., 2023; Abdul-Kader and Woods, 2015; Kim et al., 2021). Despite the powerful capabilities of large language models (LLMs), traditional SDS remain a focal point of research due to their superior control and interpretability (Singh et al., 2024). These systems are predominantly trained using data from human-to-human interactions, which highlight varying speaking styles, such as clarity of intentions, as depicted in Figure 1. This variability necessitates that human speakers adjust their dialogue strategies when engaging with different users. For instance, compared to adults, minors often exhibit less clarity in their intentions and give ambiguous responses, requiring more confirmatory language or additional inquiries to better adapt to the unique speaking styles of younger users. This adaptive approach is crucial for enhancing



Speaking Styles of User A

- There may be a clear direction intention regarding destinations and activities.
- There can be a demand for more detailed information or specific proposals.

Speaker : [DirectionQuestion] So, do you have any plans for a trip? I mean, is there a particular place you want to visit?

User A : Yes, I'd like to go to Hokkaido.

Speaker : [SeasonQuestion] Oh, I see. Do you have a preferred season for your trip?

User A : I'm thinking about going in the autumn.


Speaker : [PeopleQuestion] How many people will be traveling with you?

User A : Oh, it's just me.

Speaker : [RequestQuestion] Okay, is there anything specific you want to do or see?

User A : Yes. Well, I'd like to visit a place where I can see beautiful autumn leaves.

Speaker : [RequestConfirm, SearchInform] Alright. Let me look up some places with beautiful autumn foliage for you.



Speaking Styles of User B

- There may be a lack of clear intention regarding destinations and activities.
- Responses to questions and suggestions can sometimes be limited.

Speaker : [DirectionQuestion] Now, have you decided on a specific place you'd like to visit?

User B : Yes.

Speaker : [DirectionQuestion] Where would you like to go?

User B : Hiroshima.

Speaker : [RequestConfirm, DirectionQuestion] Hiroshima, got it. Do you have a specific area in Hiroshima in mind?

User B : No, I haven't decided yet.

Speaker : [RequestQuestion] Okay. Is there anything in particular you'd like to do there? Any activities or food you want to try?

User B : I want to visit my grandparents.

Speaker : [RequestConfirm] Oh, your grandparents. So, they live around that area.

Figure 1: Real human-to-human conversations. Speakers adopt various dialogue acts when interacting with users employing diverse speaking styles.

ing the effectiveness and user-friendliness of SDS in real-world scenarios.

However, adapting SDSs to these distinctive speaking styles typically requires a wealth of annotated dialogue data, which can be challenging to

obtain in abundance owing to the minority status of users employing unique conversational strategies or behaviors. To enhance the performance of dialogue systems when engaging with data-scarce user groups exhibiting distinct speaking styles, targeted data augmentation is imperative, enabling the system to better cater to their needs.

This study introduces a tailored data augmentation framework designed specifically for low-resource user groups exhibiting distinctive conversational behaviors. Recognizing the unique conversational behaviors and challenges associated with minors and the inherent difficulty in obtaining their data (Aydin et al., 2021), our study conducts experiments utilizing dialogue data from minors to facilitate targeted data augmentation for this demographic.

As depicted in Figure 1, the unique speaking style of users directly influences the speaker’s dialogue acts (DAs) and indirectly shape response content. Therefore, our data augmentation framework focuses on the speaking styles of users and the trajectory of DAs.

Specifically, we utilized a LLM to extract the speaking styles of such users and speakers interacting with them. We then fine-tuned a pre-trained language model (PLM) using all available data in a low-resource setting to create varied histories of DAs for speakers interacting with these user groups. The resulting speaker styles and DA histories were input into the LLM to produce customized training dialogue data for these users. The primary goal is to enhance the model’s ability to predict DAs when interacting with low-resource groups with unique speaking styles, as controlling the content of generated responses through DAs is deemed effective (Kawano et al., 2021).

This study’s contributions are outlined below.

- We introduced a data augmentation method to enhance the performance of the DA prediction model when dealing with users who have limited data and unique conversational behaviors and styles.
- Through multiple experiments conducted in a low-resource setting, we have discovered that the difficulty of DA prediction varies across different users and demonstrated the adaptability and effectiveness of our proposed method.

2 Related Work

The scarcity of annotated data and the challenge of data imbalance are persistent issues in various artificial intelligence domains (Shorten and Khoshgoftaar, 2019; Shi et al., 2020; Ahmad et al., 2021; Hedderich et al., 2021; Kim et al., 2023). To address these effectively, data augmentation techniques have been employed, as demonstrated in prior research across different tasks (Feng et al., 2021; Bayer et al., 2022). For instance, Schick and Schütze (2021) generated text similarity datasets from scratch by instructing a large PLM. Similarly, Liu et al. (2022) and Chen and Yang (2021) enhanced data by manipulating individual utterances within dialogues—such as adding, deleting, changing their order, or regenerating them—while preserving the original meaning, which improved model performance in dialogue summarization tasks. While the abovementioned methods focus on generating individual sentences, our study aims to create coherent dialogues comprising multiple sentences tailored for specific target groups.

Mohapatra et al. (2021) utilized GPT-2 (Radford et al., 2019) to develop user and agent bots, generating comprehensive task-oriented dialogues through bot interactions, demonstrating notable enhancements in low-resource scenarios with datasets MultiWOZ (Budzianowski et al., 2018) and PersonaChat (Zhang et al., 2018). Recently, with the advanced text generation capabilities of LLMs, researchers have started using LLMs for data augmentation (Pan et al., 2023; Kim et al., 2023; Wang et al., 2023). For instance, Kim et al. (2023) guided LLMs to generate a broad spectrum of social dialogues using social commonsense knowledge from a knowledge graph. Pan et al. (2023) generated domain-specific, task-oriented dialogues by extracting dialogue paths from out-of-domain conversations. The concept of dialogue paths in their work aligns with the concept of DA history in our research. However, the key distinction is that while they extract DA paths from existing data, we generate tailored DA histories based on existing data, specifically optimized for target user groups.

3 The Proposed Framework

In this study, we aim to enhance the DA prediction performance of the system when dealing with low-resource user groups that exhibit unique dialogue strategies, by generating training data through the proposed data augmentation framework. In the

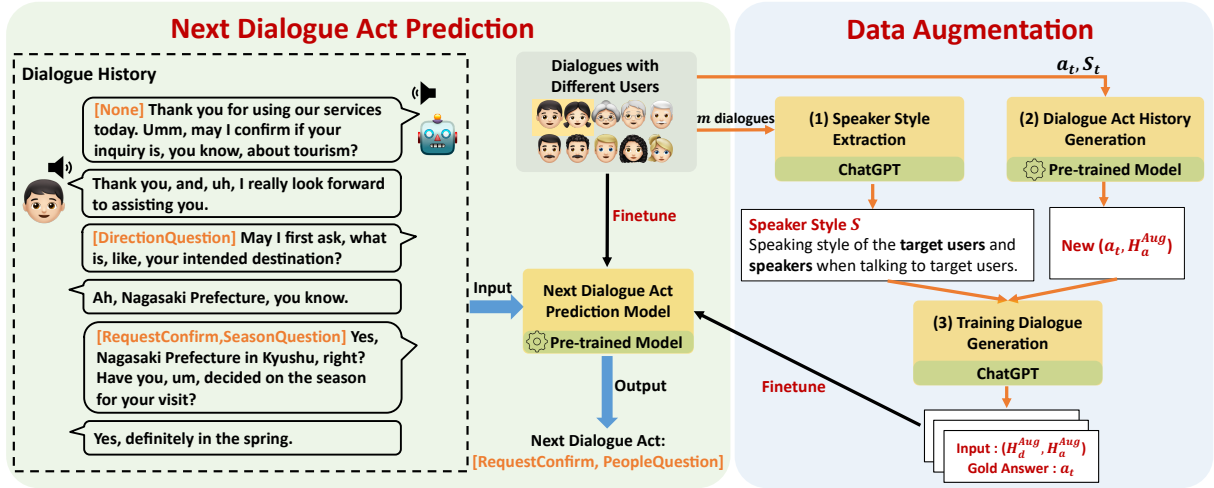


Figure 2: Our data augmentation framework is designed to improve the performance of the PLM in predicting DA when interacting with low-resource users who exhibit unique speaking styles. Beginning with dialogues that involve specific target users, we: (1) extract speaker styles, (2) generate DA histories of system interactions with these users, and (3) input this information into ChatGPT for tailored data augmentation.

construction of SDSs, accurate DA prediction is crucial as it facilitates dialogue state tracking and guides response generation, thereby reducing erroneous responses (Chen et al., 2017). The task depicted in the left portion of Figure 2 is defined as follows. Assuming the current turn of the dialogue is turn t , we utilize the dialogue history $H_d = (S_{t-n}, U_{t-n}, \dots, S_{t-1}, U_{t-1})$ from the previous n turns, along with the system’s DA history $H_a = (a_{t-n}, \dots, a_{t-1})$ from these turns, as the input. The output is the system’s DA a_t for the current turn.

Since we predict the current turn’s DA based on the dialogue history and the system’s DA history, it becomes crucial to generate dialogue and system DA histories that closely align with the target user group. To achieve this, we control the generation of dialogue data by capturing the speaking style of dialogue participants and generating dialogue flows that mimic real human interactions with the target user group. The importance of this approach lies in the fact that the model can effectively understand and adapt to unique dialogue strategies only when the training data realistically simulates complex dialogue scenarios. In real human interactions, users with unique dialogue strategies are in the minority and exhibit considerable diversity. Due to the limitations in data scale, traditional training datasets often fail to cover this diversity, which limits the model’s adaptability and accuracy when dealing with such users. By simulating the dialogue styles and processes of specific user groups, we can gener-

ate more diverse and precise training data, thereby enhancing the model’s generalizability and adaptability to diverse users.

As illustrated in Figure 2, our data augmentation framework comprises three components: (1) employing ChatGPT¹ to extract the speaker’s styles S , (2) finetuning a pre-trained model to generate the system’s DA history $H_a^{Aug} = (a_{t-n}^{Aug}, \dots, a_{t-1}^{Aug})$, and (3) inputting the extracted speaking styles S and the generated system’s DA history H_a^{Aug} into ChatGPT to generate the training dialogue data $H_d^{Aug} = (S_{t-n}^{Aug}, U_{t-n}^{Aug}, \dots, S_{t-1}^{Aug}, U_{t-1}^{Aug})$.

3.1 Speaker Styles Extraction

Since the unique speaking styles employed by the target user group significantly influence the content of conversations, it’s crucial to capture the speaking styles of this group by comparing dialogues from the target user group with those from non-target groups. This helps guide the subsequent generation of dialogues specifically tailored to the target user group. To facilitate this, we employ ChatGPT to extract speaker styles from conversations involving target users.

Specifically, we input a set of m dialogues, half of which involve users from the target group and the other half from non-target user groups. This balanced approach allows for an effective comparison, helping to identify and differentiate prominent speaking characteristics unique to the target group. Subsequently, ChatGPT is utilized to generate out-

¹<https://openai.com/blog/chatgpt>

puts representing the speaking style of the target users, as well as the speaking style of speakers when engaging with the target user group. Notably, our primary focus is on extracting abstract styles, such as "target users often exhibit ambiguous intentions towards destinations and activities." These styles are crucial because they significantly influence the direction of the dialogue, thereby enhancing the realism and relevance of the generated dialogues to actual human conversations. The prompt and extracted speaker styles are presented in Appendix C.

3.2 DA History Generation

As depicted in Figure 1, the unique conversational strategies employed by the target group also significantly influence the DAs of those engaging with them. Our objective at this stage is to generate a diverse and realistic DA history H_a^{Aug} that is specifically optimized for groups with distinctive speaking strategies. As shown in Figure 3, we achieve this by finetuning a PLM using existing data to generate the system’s DA history H_a^{Aug} for the previous n turns.

In particular, we utilize the DA a_t and utterance S_t from the current turn t as inputs, with the DA history H_a from the previous n turns as the desired output to establish training data. These data are then divided into two sets: one for training and the other for generation. Initially, we finetune the PLM using all available training data to capture DA histories that closely resemble real human conversations. Subsequently, we conduct a secondary finetuning utilizing training data exclusively from the target user group. This dual finetuning approach ensures that the model can generate DA histories that closely mimic real human dialogues and align with the unique speaking strategies of the target users. The first finetuning, which employs a relatively large dataset, enables the model to produce DA histories that mirror authentic human interactions. The second finetuning, focused on a smaller dataset specific to the target user group, allows the model to better tailor the DA histories to their unique characteristics.

During the generation phase, we input the the DA a_t and utterance S_t from the current turn t and generate the DA history H_a^{Aug} from the previous n turns. To ensure diversity, we simultaneously generate multiple outputs, selecting only those (a_t, H_a^{Aug}) combinations that have not been previously observed.

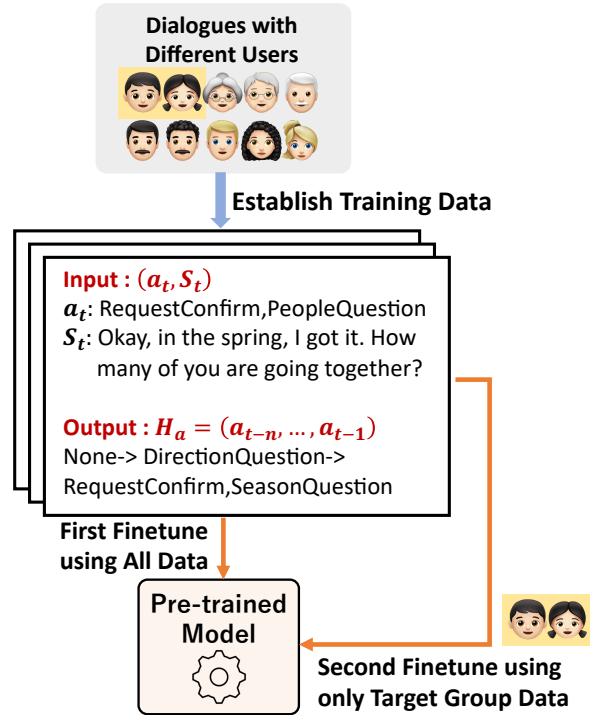


Figure 3: DA History Generation. We conduct two rounds of finetuning: the first round using all available data, and the second round using only data from the target user group, to ensure the generated DA history more closely aligns with the target demographic.

3.3 Dialogue Generation

Having obtained speaker styles and DA history tailored to users employing unique dialogue strategies, our ultimate goal is to generate dialogues corresponding to these styles and histories to enrich the training data for DA prediction. At this stage, we leverage ChatGPT’s powerful generation capabilities to create dialogue data for training purposes. Utilizing a few-shot prompt, we input the extracted speaking styles S and the DA histories H_a^{Aug} into ChatGPT to generate dialogues H_d^{Aug} that reflect the conversational style of the target users. Subsequently, we use the generated dialogues H_d^{Aug} and DA histories H_a^{Aug} as inputs, with a_t as the gold-standard answer, to construct the training data. The prompts used for generating these dialogues are detailed in Appendix D.

This approach aims to enhance the model’s ability to predict DAs when interacting with target users who exhibit unique conversational strategies. It effectively addresses the challenge of data scarcity by employing data augmentation.

4 Experiment

To evaluate the effectiveness of the proposed data augmentation framework, we conducted experiments using data from minors who employed unique conversational styles and strategies in actual dialogues within the dataset. These experiments were carried out in a low-resource setting across multiple splits, each utilizing different subsets of data from minors. We trained multiple DA prediction models on datasets of varying sizes, including models trained with augmented data added to the existing datasets.

4.1 Dataset

This study utilized a multimodal dialogue Japanese dataset known as the “Travel Agency Task Dialogue Corpus” (Inaba et al., 2022, 2024), which features conversations from users of various age groups, with detailed annotations of DAs. This dataset contains 115 hours of dialogue, spanning 330 conversations, with each averaging about 20 minutes. The dialogues were facilitated via Zoom video calls, involving six operators and 55 customers, including 20 minors (ages 7-17), 25 adults (ages 20-60), and 10 seniors (ages 65-72). Each customer participated in six dialogues.

The dialogues revolve around recommending travel destinations to users across various age groups. The dataset authors employed a Hidden Markov Model (HMM) (Rabiner, 1989) to analyze the transitions in dialogue among different age groups using sequences of DAs. A notable observation was that minors often used unique dialogue strategies compared to other age groups, typically expressing fewer independent opinions. The annotation of DAs was performed by functional segment, a unit smaller than an utterance. Each operator’s segment is annotated as one of the 28 predefined DAs related to travel destination recommendations, or as “None”. Examples of these DAs include asking about the travel season (Season-Question) and summarizing the travel plan (Travel-Summary), all of which are detailed in Appendix A. Since segments labeled “None” primarily consist of non-informative responses such as “Yeah” or “Uh-huh,” and our objective is to guide the system to generate accurate and meaningful responses using DA tags, we selectively included only those training instances where the gold-standard responses were not labeled “None” in this study. Additionally, we employed text-based human transcriptions

rather than audio recordings for our research.

4.2 Low-Resource Setting

We trained five DA prediction models using datasets of varying scales: Minors-Only, Zero-Shot, Low-Resource, Full-Resource, and Low-Resource+Augmentation(Ours). To simulate low-resource conditions for specific user demographics, we used dialogue data from only 3 minors out of a group of 20, totaling 18 dialogues for training. For evaluation, we used 60 dialogues from 10 minors.

- **Minors-Only:** Employed only 18 dialogues from 3 minors.
- **Zero-Shot:** Utilized all data from adults and seniors, amounting to 210 dialogues.
- **Low-Resource:** Combined the 18 dialogues from the Minors-Only with all 210 dialogues from adults and seniors, totaling 228 dialogues.
- **Full-Resource:** Included dialogues from 10 minors (60 dialogues), encompassing those from the 3 minors in the low-resource setting, plus all 210 dialogues from adults and seniors, totaling 270 dialogues.
- **Low-Resource + Aug(mentation) (Ours):** Used the 228 dialogues from the Low-Resource and supplemented them using our proposed augmentation framework. Additional data was generated until the dataset size matched that of the Full-Resource for a direct comparison.

4.3 Setup and Details

In the process of extracting speaker styles, we fed $m = 6$ dialogues into GPT-4-0125-preview, where three were from minors in a low-resource setting, and the other three involved different adults or seniors. For generating training dialogues, GPT-3.5-turbo-0125 was employed.

During the DA history generation phase, we utilized Japanese T5-Large² as the PLM. We conducted two rounds of finetuning to ensure the model is capable of generating DA histories that not only closely mimic real human conversations but also align with the unique conversational strategies of minors during interactions. During the first training phase, the learning rate was set at 1e-4, and

²<https://huggingface.co/retrieva-jp/t5-large-long>

Table 1: Training data quantity for DA prediction across four splits: MO (Minors-Only), ZS (Zero-Shot), LR (Low-Resource), FR (Full-Resource)

Split	Valid	MO-Valid	MO	ZS	LR	FR	Ours	Test
1	2,027	307	1,662	21,011	22,980	26,375	26,375	6,004
2	2,027	199	1,117	21,011	22,327	26,434	26,434	5,945
3	2,027	262	1,578	21,011	22,851	26,712	26,712	5,667
4	2,027	271	1,574	21,011	22,856	26,961	26,961	5,418

for the subsequent phase exclusively involving data from minors, it was set at $5e-5$. We utilized 210 adult and elderly conversations for generating DA histories, dividing them into 120 for training and 90 for generation purposes. To ensure data diversity and novelty, we retained only those (a_t, H_a^{Aug}) combinations that had not previously existed; all 18 dialogues from 3 minors were included in both training and generation phases. To ensure diversity, we set the `num_return_sequences=3` when generating DA histories, meaning that for each data point, three DA histories are generated simultaneously.

In the DA prediction phase, Japanese T5-base³ and Japanese GPT-NeoX⁴ were used as the PLMs to validate the effectiveness of the generated data. We reconstructed the training and evaluation sets for the same DA prediction task to optimize hyperparameters, with specific details provided in Appendix B. Regarding the distribution of training and validation sets, the validation sets for all settings, except Minors-Only, are identical, comprising 21 dialogues from adults and seniors. The Minors-Only validation set consists of 3 dialogues from minors in the low-resource scenario. To validate the generalizability of our method, we conducted experiments across four splits, each using data from three different minors for training under a low-resource setting, while also varying the test data. Details on the data points for each split, after removing entries with a gold-standard answer of "None," are outlined in Table 1.

Considering that a single utterance may consist of multiple segments (see Figure 1 and Figure 2), each potentially be labeled with a different DA, there may be more than one gold-standard DA label for the current turn. Therefore, we employed both **exact match and partial match rates** as evaluation metrics. The exact match rate is a strict metric requiring the predicted set of labels to completely align with the true set of gold labels, measuring the model’s ability to fully grasp the dialogue con-

text and predict all relevant DA labels accurately. The partial match rate assesses the model’s performance in predicting some correct labels. This metric is more lenient, recognizing that in real conversations, capturing the main intent or action of the dialogue, even if not every label is precisely predicted, is still valuable. Therefore, the partial match rate helps understand the model’s robustness in practical use. Combined, these two metrics offer a balanced approach to evaluating the model’s DA prediction capabilities, providing a more accurate reflection of the model’s performance.

5 Results and Analysis

Table 2 shows the mean and standard deviation after five runs using seeds ranging from 1 to 5 across four different splits. While the **Minors-Only** solely comprised data from minors, its performance was inferior to the **Zero-Shot** model trained only with adult and elderly dialogue data due to the limited amount of training data. Therefore, we also used all available adult and elderly dialogue data in other setups to enhance the model’s generalization capabilities.

Additionally, since **Zero-Shot** does not use minor’s dialogues, the training data remains consistent across the four different splits. The variation in **Zero-Shot**’s performance across the splits further underscores the differences in the model’s adaptability to different minors, with the third split proving most challenging.

Across the four splits, the performance of our proposed data augmentation framework, **Low-Resource + Aug (Ours)**, almost all surpassed that of **Low-Resource** on both T5 and GPT-NeoX in terms of mean exact and partial match rates. This demonstrates that even in a low-resource setting, our method successfully captures the characteristics of minor speakers and generates dialogue flows that align with minor speaking behaviors, thereby guiding the generation of training dialogues.

However, even though we augmented the data to match the quantity of the **Full-Resource** in each split, **Full-Resource** typically showed superior per-

³<https://huggingface.co/retrieva-jp/t5-base-long>

⁴<https://huggingface.co/stockmark/gpt-neox-japanese-1.4b>

Table 2: Results across four different splits.

Split	Setting	Japanese GPT-NeoX		Japanese T5-base	
		Exact Match	Partial Match	Exact Match	Partial Match
1	Minors-Only	0.2451 ± 0.0117	0.3447 ± 0.0131	0.2533 ± 0.0083	0.3519 ± 0.0090
	Zero-Shot	0.2966 ± 0.0071	0.4049 ± 0.0092	0.3000 ± 0.0059	0.4066 ± 0.0053
	Low-Resource	0.3041 ± 0.0070	0.4228 ± 0.0073	0.3085 ± 0.0065	0.4232 ± 0.0064
	Low-Resource + Aug (Ours)	0.3137 ± 0.0064	0.4320 ± 0.0094	0.3148 ± 0.0050	0.4244 ± 0.0056
	Full-Resource	0.3190 ± 0.0074	0.4489 ± 0.0049	0.3125 ± 0.0029	0.4418 ± 0.0023
2	Minors-Only	0.2302 ± 0.0103	0.3677 ± 0.0105	0.2419 ± 0.0050	0.3311 ± 0.0079
	Zero-Shot	0.3162 ± 0.0069	0.4247 ± 0.0099	0.3200 ± 0.0039	0.4263 ± 0.0046
	Low-Resource	0.3220 ± 0.0071	0.4401 ± 0.0051	0.3257 ± 0.0019	0.4430 ± 0.0066
	Low-Resource + Aug (Ours)	0.3290 ± 0.0083	0.4460 ± 0.0111	0.3270 ± 0.0029	0.4473 ± 0.0095
	Full-Resource	0.3294 ± 0.0068	0.4526 ± 0.0074	0.3339 ± 0.0052	0.4486 ± 0.0075
3	Minors-Only	0.2329 ± 0.0033	0.3291 ± 0.0069	0.2528 ± 0.0038	0.3499 ± 0.0010
	Zero-Shot	0.2771 ± 0.0053	0.3878 ± 0.0075	0.2787 ± 0.0054	0.3889 ± 0.0054
	Low-Resource	0.2863 ± 0.0055	0.4070 ± 0.0019	0.2825 ± 0.0036	0.4010 ± 0.0156
	Low-Resource + Aug (Ours)	0.2906 ± 0.0055	0.4077 ± 0.0067	0.2865 ± 0.0042	0.4097 ± 0.0090
	Full-Resource	0.2889 ± 0.0069	0.4282 ± 0.0085	0.2986 ± 0.0058	0.4270 ± 0.0057
4	Minors-Only	0.2325 ± 0.0083	0.3336 ± 0.0093	0.2429 ± 0.0036	0.3480 ± 0.0091
	Zero-Shot	0.2900 ± 0.0066	0.4041 ± 0.0066	0.2947 ± 0.0047	0.4056 ± 0.0059
	Low-Resource	0.2925 ± 0.0067	0.4098 ± 0.0088	0.2983 ± 0.0031	0.4156 ± 0.0120
	Low-Resource + Aug (Ours)	0.3005 ± 0.0069	0.4254 ± 0.0087	0.3000 ± 0.0056	0.4144 ± 0.0096
	Full-Resource	0.3096 ± 0.0049	0.4425 ± 0.0098	0.3094 ± 0.0073	0.4336 ± 0.0019

formance. A possible explanation is the lack of quality control, which meant that subpar data was not filtered out, leading to poorer adaptation compared to **Full-Resource**, which used data exclusively from real human conversations. Additionally, the "Travel Agency Task Dialogue Corpus," derived from video calls and manually transcribed, may contain colloquial filler words and other informal elements in its complete utterances. In contrast, ChatGPT-generated dialogues tend to be more structured and fluid. This stylistic difference could also contribute to the observed performance disparity between **Low-Resource + Aug (Ours)** and **Full-Resource**.

5.1 Ablation

To evaluate the individual effectiveness of components in our proposed framework, we conducted ablation experiments using Japanese GPT-NeoX across four splits:

- **w/o DA History Gen:** In this model, we omitted the generation of new DA histories and instead randomly selected DA histories from the Low-Resource for data generation.
- **DA History Gen w/o Second Finetune:** This variant involved finetuning the DA history generation model only once, without a second round of finetuning tailored specifically for minors.
- **w/o Speaker Style:** This model utilized the

same DA histories as our complete method but did not use extracted speaker styles during dialogue data generation.

Table 3 shows the average results across the four splits, conducting five trainings for each model in every split with seed values set from 1 to 5. The findings indicate that both **w/o DA History Gen** and **w/o Speaker Style** variants achieved higher mean exact and partial match rates than the **Low-Resource**. This demonstrates that the training data generated through the independent use of style extraction and DA history generation components can also significantly improve performance.

Furthermore, although **DA History Gen w/o Second Finetune** did not use data from the target user group for a second fine-tuning during the training of the DA history generation model, its performance still surpassed that of **w/o DA History Gen**. This indicates that in generating DA history, even without a second finetuning to optimize the PLM for minors, the new (a_t, H_a^{Aug}) combinations generated by a PLM trained with all available data can still enhance performance. Ultimately, **Ours** achieved the highest rates for both exact and partial matches, indicating that the combination of speaker styles extraction and DA history generation is most effective and underscores the necessity of targeted age-specific second finetuning when training the DA history generation model.

Table 3: Average results of the ablation experiments across four splits.

Setting	Exact Match	Partial Match
Low-Resource	0.3012	0.4199
w/o DA History Gen	0.3052	0.4263
DA History Gen w/o Second Finetune	0.3072	0.4269
w/o Speaker Style	0.3027	0.4274
Ours	0.3085	0.4278

Speaker Style

Speaking style of the target users:

- Tends to lack a concrete image of the destination or activities, using vague expressions.
- More likely to have sensory-based desires rather than specific plans. For example, "want to enjoy nature" or "prefer a quiet place with few people". (...)

Speaking style of the interlocutors when conversing with target users:

- Make suggestions while checking the speaker's response.
- Ask questions to draw out the speaker's interests and desires. (...)

Dialogue Act History

RequestQuestion -> RequestConfirm -> SearchConditionInform

w/o Speaker Style

System: 旅行先では何を楽しみにされてますか？(What are you looking forward to when traveling?)

User: あの、お寺や神社が見たいです。(Oh, I want to see temples and shrines.)

System: 了解しました。お寺や神社の観光地をお探しいたしますね。(Understood. I'll look for tourist spots with temples and shrines for you.)

User: よろしくお願ひします。(Thank you.)

System: では、京都の清水寺や金閣寺など、たくさんのお寺や神社があるエリアがおすすめです。どういたしまして。(In that case, I recommend areas like Kyoto, where there are many temples and shrines, such as Kiyomizu Temple and Kinkaku-ji. You're welcome.)

User: はい。(Yes.)

Ours

System: 旅行の際に特にこだわりや希望はありますか？(Do you have any particular preferences or desires for your trip?)

User: うーん、特にないです。(Hmm, not really.)

System: そうですね、何も特にこだわりがないということですね。(I see, no specific preferences then.)

User: はい。(Yes.)

System: その場合、近場で穏やかな雰囲気が楽しめる場所をおすすめします。どうでしょうか？(In that case, I recommend somewhere nearby with a calm atmosphere. How does that sound?)

User: いいですね。(That sounds nice.)

Figure 4: Dialogues generated by the variant without speaker styles and our approach.

5.2 Why did the Speaker Style work?

Figure 4 displays dialogues generated by **w/o Speaker Style** and **Ours**, using the same DA history. The DA history consists of first asking the user a travel-related request (RequestQuestion), then confirming the request (RequestConfirm), and finally indicating the content to be searched (SearchConditionInform). We observed that without the speaker style, the user in the **w/o Speaker Style** provided specific travel requirements, and the dialogue progressed smoothly. In contrast, the user in the **Ours** did not exhibit a clear intent. This indicates that the speaker style is effective, resulting in dialogues that more closely match the speaking styles of minors and aligning more closely with real human conversations.

5.3 Why did the DA History Generation work?

We compared the performance in generating DA histories between **DA History Gen w/o Second**

Finetune and Ours on split 1.

For a direct comparison, we used 9,999 data points (a_t, S_t) from dialogues involving 90 adults and seniors to generate DA histories H_a^{Aug} , resulting in three DA histories per data point. This generation was conducted under the settings of $top_k=50$, $top_p=0.9$, and $temperature=0.9$. After removing duplicate (a_t, H_a^{Aug}) , **DA History Gen w/o Second Finetune** produced 7,677 new (a_t, H_a^{Aug}) , whereas **Ours** generated 10,412. We assessed how many of these combinations appeared in dialogues involving 17 minors (excluding those from the **Low-Resource**), finding 908 for **DA History Gen w/o Second Finetune** and 956 for **Ours**. Referencing Table 3, we can infer that compared to **w/o DA History Gen** which relied solely on existing DA histories, both **DA History Gen w/o Second Finetune** and **Ours** generated DAs that were present in the target user group, leading to improved performance. Notably, **Ours**, which underwent secondary finetuning for the target users,

produced more DA histories closely aligned with the target group, enhancing performance.

6 Conclusion

We introduced a data augmentation method designed to enhance the performance of the DA prediction model for users with limited data and unique conversational styles. Our experiments confirmed the reliability of the proposed method and the effectiveness of its components. While this study did not exhaustively explore the full potential for improvement of the proposed method, we plan to further evaluate this aspect in our future work.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19H05692.

References

- Sameera A. Abdul-Kader and Dr. John Woods. 2015. [Survey on chatbot design techniques in speech conversation systems](#). *International Journal of Advanced Computer Science and Applications*, 6(7).
- Tanveer Ahmad, Dongdong Zhang, Chao Huang, Hongcai Zhang, Ningyi Dai, Yonghua Song, and Huanxin Chen. 2021. [Artificial intelligence in sustainable energy industry: Status quo, challenges and opportunities](#). *Journal of Cleaner Production*, 289:125834.
- Selami Aydin, Leyla Harputlu, Özgehan Uştuk, Şeyda Savran Çelik, and Serhat Güzel. 2021. Difficulties in collecting data from children aged 7–12. *International Journal of Teacher Education and Professional Development (IJTEPD)*, 4(1):89–101.
- Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. [A survey on data augmentation for text classification](#). *ACM Comput. Surv.*, 55(7).
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. [MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.
- Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. [A survey on dialogue systems: Recent advances and new frontiers](#). *SIGKDD Explor. Newsl.*, 19(2):25–35.
- Jiaao Chen and Diyi Yang. 2021. [Simple conversational data augmentation for semi-supervised abstractive dialogue summarization](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. [A survey of data augmentation approaches for NLP](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.
- Michael A. Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. [A survey on recent approaches for natural language processing in low-resource scenarios](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.
- Michimasa Inaba, Yuya Chiba, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. 2022. [Collection and analysis of travel agency task dialogues with age-diverse speakers](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5759–5767, Marseille, France. European Language Resources Association.
- Michimasa Inaba, Yuya Chiba, Zhiyang Qi, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. 2024. [Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*
- Tatsuya Kawahara. 2018. Spoken dialogue system for a human-like conversational robot erica. In *International Workshop on Spoken Dialogue Systems Technology*.
- Seiya Kawano, Koichiro Yoshino, and Satoshi Nakamura. 2021. [Controlled neural response generation by given dialogue acts based on label-aware adversarial learning](#). *Transactions of the Japanese Society for Artificial Intelligence*, 36(4):E-KC9₁ – –14.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. [SODA: Million-scale dialogue distillation with social commonsense contextualization](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.
- Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Z. Hakkani-Tür. 2021. How robust r u?: Evaluating task-oriented dialogue systems on spoken conversations. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154.
- Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. 2022. [Data augmentation for low-resource dialogue summarization](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages

- 703–710, Seattle, United States. Association for Computational Linguistics.
- Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2021. [Simulated chats for building dialog systems: Learning to generate conversations from instructions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1190–1203, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yan Pan, Davide Cadamuro, and Georg Groh. 2023. Data-augmented task-oriented dialogue response generation with domain adaptation. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 96–106, Hong Kong, China. Association for Computational Linguistics.
- L.R. Rabiner. 1989. [A tutorial on hidden markov models and selected applications in speech recognition](#). *Proceedings of the IEEE*, 77(2):257–286.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Timo Schick and Hinrich Schütze. 2021. [Generating datasets with pretrained language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Wenzhong Shi, Min Zhang, Rui Zhang, Shanxiong Chen, and Zhao Zhan. 2020. [Change detection based on artificial intelligence: State-of-the-art and challenges](#). *Remote Sensing*, 12(10).
- Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48.
- Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. [SpokenWOZ: A large-scale speech-text benchmark for spoken task-oriented dialogue agents](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. [Rethinking interpretability in the era of large language models](#).
- Xi Wang, Hossein Rahmani, Jiqun Liu, and Emine Yilmaz. 2023. [Improving conversational recommendation systems via bias analysis and language-model-enhanced data augmentation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3609–3622, Singapore. Association for Computational Linguistics.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting*
- of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A DA tags in Travel Agency Task Dialogue Corpus

In this study, we utilized the "Travel Agency Task Dialogue Corpus" collected by Inaba et al. (2024), which includes task specific DA annotations. The dataset defines DA tags for operators and customers in travel agency conversations, with 28 tags for operators and 8 tags for customers. In this study, only the operator’s tags were used, as shown in Table 4.

Table 4: Task Specific Dialogue Act Tags for Operator Segments.

Dialogue Act	Description	Example
DirectionQuestion	Question on areas for the desired travel	To which destination are you planning to travel?
SeasonQuestion	Question on the desired season	When will you go?
PeopleQuestion	Question about the number of people traveling and their relationships with the customer	How many people are traveling with you?
AgeQuestion	Question on the age of customers or their companions	How old are your children?
ExperienceQuestion	Question about the customer’s experience	Have you ever been to Osaka?
RequestQuestion	Question about the tourist spot request	What would you like to do there?
SearchAdvice	Questions or suggestions related to the tourist spot information retrieval system	Should I look for a restaurant there?
RequestConfirm	Confirmation of requests for tourist spots	You want to go to a Spa, don’t you?
DestinationConfirm	Confirmation of destination	Am I correct in Assuming that you are going to Yashi Park?
AddDestinationList	Addition to destination list by operator	I’ll add this location to the list.
TravelSummary	Summary of trip planning	Looking back, you plan to visit the Toshogu Shrine first.
SearchInform	Operator’s declaration of intent to search tourist spots in the system	I will now search.
PhotoInform	Provide information on photos displayed on the system	Here is a picture of a meal containing a lot of salmon roe.
SearchConditionInform	Provide information on search conditions	I can also filter by the time required.
NameInform	Provide information on the names of tourist spots	There is a commercial complex called the Sapporo Factory.
IntroductionInform	Provide information on tourist spots based on the system search results	It was established In 1876.
OfficeHoursInform	Provide information on hours of operation and closing dates	Our business hours span 10:00 a.m. to 10:00 p.m.
PriceInform	Provide information on fees and price range	The admission fee is 360 yen.
FeatureInform	Providing information about the characteristics of tourist spots	It is recommended for women even when it rains.
AccessInform	Provide information on access	This location is a five-minute walk from the railway station.
PhoneNumberInform	Provide information on telephone numbers	The phone number is 095 824.
ParkInform	Provide information on parking	There are three parking lots.
EmptyInform	Statement that there are no search results or specific description	I do not see anything in the search results.
MistakeInform	Correcting errors in tourist spot information	Sorry, this store is open on all days of the week.
OperatorSpotImpression	Subjective evaluations and assumptions about a tourist spot by operators	This restaurant looks nice and inexpensive.
SearchResultInform	Report overall search results	It appears there are numerous stores in this location.
OnScreenSuggest	Suggestions for tourist spots on the shared screen	How about this site?
OnScreenQuestion	Questions about tourist spots on the shared screen	Which one looks the best, number 1, 2, or 3?

B Hyperparameter Optimization

During our experiments, we performed hyperparameter optimization.

For T5-base, we conducted a grid search with batch sizes of {8, 16, 32, 64}, warmup ratios of {0, 0.1, 0.2}, and learning rates of {3e-3, 2e-3, 1e-3, 9e-4, 8e-4}. The optimal configuration was identified as a batch size of 64, a warmup ratio of 0.1, and a learning rate of 1e-3.

Similarly, for GPT-NeoX, we conducted a grid search with batch sizes of {4, 8, 16}, warmup ratios of {0.1, 0.2, 0.3}, and a range of learning rates of {3e-4, 2e-4, 1e-4, 9e-5, 8e-5, 7e-5, 6e-5, 5e-5, 4e-5}. The best settings were determined to be a batch size of 8, a warmup ratio of 0.1, and a learning rate of 9e-5.

C Details for Speaker Styles Extraction.

We utilized the prompt shown in Figure 5 to extract speaker styles using the GPT-4-0125-preview model, with six dialogues from different users, three from the target user group and three from a non-target user group. As the extraction was conducted with the default temperature setting (i.e., temperature=1), the generated results were diverse. We performed multiple extractions and manually combined the extracted speaker styles. The consolidated speaker styles, as illustrated in Figure 6, were all used for subsequent dialogue data generation.

```
# Task Description
The task involves providing tourist destination guidance in dialogues for three minor users and three general users. The objective is to summarize the styles of speakers in the target age group and the speaking styles of the speakers interacting with them in comparison to the given dialogues. Please outline these in bullet points, detailing as much as possible.

# Target Age Group Dialogue 1
Speaker: [RequestQuestion] May I ask about your travel plans?
User: Well, I'm thinking of going to Okinawa in the spring.
Speaker: [RequestConfirm] Spring in Okinawa, right?
User: Yes.
Speaker: [DirectionQuestion] Do you have a specific area in Okinawa in mind?
User: Not really, I haven't decided yet.
(...)

# Target Age Group Dialogue 2
(...)

# Target Age Group Dialogue 3
(...)

# Non-target Age Group Dialogue 1
(...)

# Non-target Age Group Dialogue 2
(...)

# Non-target Age Group Dialogue 3
(...)

# Answer
```

Figure 5: Prompt for Speaker Styles Extraction.

```
# Speaker Style S
Speaking style of the target users:

- Tends to lack a concrete image of the destination or activities, using vague expressions.
- More likely to have sensory-based desires rather than specific plans. For example, "want to enjoy nature" or "prefer a quiet place with few people."
- They often express general hopes rather than detailed plans.
- They often speak while thinking, using phrases like "umm" or "well."
- They frequently respond with just "yes."
- Their statements can be short, hesitant, and sometimes unclear in meaning.
- They are not very knowledgeable about tourist spot names or geographical locations.
- They might give vague answers about food preferences (e.g., "I like meat, but seafood sounds good too").

Speaking style of the interlocutors when conversing with target users:

- Uses friendly and approachable words.
- Often focuses on suggesting leisure and activities, emphasizing proposals that highlight scenery and experiences.
- They strive to provide suggestions that match the minor's motivations and interests, often naming specific spots.
- They explain the features and highlights of tourist spots in detail.
- They make suggestions while checking the minor speaker's reactions.
- For minor speakers, clerks often present multiple options and encourage them to choose what interests them.
- Clerks try to understand the minor speaker's interests and needs, providing more information and asking questions to confirm.
- They ask many questions to draw out the speaker's interests and desires.
- They propose activities that might interest young speakers (e.g., interactive attractions, photo spots).
- They strive to make suggestions suitable for the season and time of day.
- They respond flexibly and make suggestions even when the speaker's requests are unclear.

```

Figure 6: Extracted Speaker Styles. They are utilized for subsequent dialogue generation.

D Prompt used for Training Dialogue Generation.

The prompt shown in Figure 7 was employed to instruct GPT-3.5-turbo-0125 to generate dialogue data for training. We included seven examples in the prompt to control the quality of generation. All examples originated from real conversations of the target user group in the "Travel Agency Task Dialogue Corpus" (Inaba et al., 2024).

```
# Task Description
Generate a travel destination recommendation dialogue from dialogue acts based on the given speaker styles.

# Speaker Style S
Speaking style of the target users:
  • Tends to lack a concrete image of the destination or activities, using vague expressions.
  • More likely to have sensory-based desires rather than specific plans. For example, "want to enjoy nature" or "prefer a quiet place with few people." (...)
Speaking style of the interlocutors when conversing with target users:
  • Uses friendly and approachable words.
  • Often focuses on suggesting leisure and activities, emphasizing proposals that highlight scenery and experiences. (...)

# Example 1
==Dialogue Act==
SeasonQuestion, RequestConfirm, PeopleQuestion
==Generated Dialogue==
System : [SeasonQuestion] Have you decided on the season for your trip?
User : I would prefer winter.
System : [RequestConfirm] Winter, I see.
User : Yes.
System : [PeopleQuestion] Understood. How many people will be traveling?
User : Well, I'd like to travel with my sister, so two of us.

# Other Examples (2~7)
(...)

# Target
==Dialogue Act==
 $a_{t-n}, \dots, a_{t-1}$ 
==Generated Dialogue==
```

Figure 7: Prompt for Dialogue Generation. Red indicates the condition generated in previous steps.