Style-Controlled Response Generation for Dialog Systems with Intimacy Interpretation

Takuto Miura, Kiyoaki Shirai, Natthawut Kertkeidkachorn
Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
{s2460005, kshirai, natt}@jaist.ac.jp

Abstract

This paper proposes a novel method to control the style of the dialog system's utterances according to the user's level of intimacy with the system. Specifically, the dialog model generates responses in a polite style when the user exhibits a low level of intimacy with the system and in a casual style when the user's intimacy is high. The proposed model consists of two submodels: the Intimacy Interpreter and Response Generator. The Intimacy Interpreter generates an embedding that represents the user's intimacy. This model is trained by contrastive learning using an intimacy-labeled dialog corpus. The Response Generator accepts a dialog context and an intimacy embedding, and then generates a response in an appropriate style. We apply two loss functions to fine-tune a Large Language Model (LLM) to train the Response Generator. The results of automatic and human evaluations show that the proposed method outperforms the baselines in terms of style control in response generation.

1 Introduction

In recent years, free dialog systems that allow users to converse about any topic have attracted considerable attention (Khatri et al., 2018; Higashinaka et al., 2021; Dinan et al., 2020). These systems need to have a comfortable conversation with the user and establish a long-term friendly relationship to facilitate conversation between the user and the dialog system (Ram et al., 2018).

To establish friendly relationships, humans change their speech style based on their level of intimacy and social connections with others to facilitate smooth communication (Wardhaugh and Fuller, 2021; Hovy, 1987; Silverstein, 2003). This ability is referred to as "style control" hereafter. The style control should also be considered in conversations between a human and a system (Kageyama et al., 2018). Consequently, a free dialog system is required to have the capability for style control.

The goal of this research is to develop a dialog system that dynamically adjusts styles according to the user's feelings toward the dialog system. A typical example of style control is that a speaker uses formal/polite expressions or informal/casual expressions by the relationship with their partner (Aapakallio, 2021; Liu and Kobayashi, 2022). Miura et al. (2024a) reported that speakers tend to use a polite style when intimacy with a partner is low and a casual style when intimacy is high. Therefore, we aim to dynamically recognize the user's level of intimacy through their dialog history and enable the dialog system to flexibly use a polite or casual style when intimacy is low or high.

This paper proposes a model that accurately identifies the user's level of intimacy with the dialog system and generates responses in an appropriate style. An intimacy interpreter is introduced to obtain a user embedding that represents the user's intimacy, and then this embedding is fed into a response generator, which is obtained by fine-tuning a Large Language Model (LLM), as a soft prompt. It enables the dialog system to appropriately control polite and casual styles.

The contributions of this paper are summarized as follows.

- We develop a dialog system that dynamically captures the user's intimacy and adjusts responses to be either polite or casual style accordingly.
- We propose a new framework to obtain an abstract representation of the user's intimacy and incorporate it into a dialog model for style control.
- The effectiveness of the proposed method is demonstrated through automatic and human evaluations.

2 Related Work

Methods for generating responses in a particular style have been actively studied. Niu and Bansal (2018) defined such a task, created a model for identifying a speech style, and proposed a method for generating responses in a given style (e.g., a polite or casual style). Gao et al. (2019) proposed a model that generated responses in a given style while maintaining consistency with the dialog context by sharing the latent space between conversational modeling and style modeling. Zhu et al. (2021) assumed that conversational modeling and style modeling are contradictory, and proposed a method to separate the representations of content and style within the shared latent space proposed by Gao et al. (2019), where each is represented in different dimensions of the latent space. Zheng et al. (2021) proposed a method for automatically constructing a dialog corpus containing utterances in a given style, which was used to train a dialog model that generated responses in line with the specified style. Specifically, they trained a Seq2Seq (Sequence-to-Sequence) model that transformed a sentence into an equivalent sentence in the specified style using a text corpus of that style. A new dialog corpus was constructed by converting the style of utterances in an original dialog corpus using the trained style conversion model. Yang et al. (2020) proposed STYLEDGPT to fine-tune a pre-trained language model to obtain a dialog model that generates utterances in the target style. They designed loss functions for fine-tuning, which were based on a language model of a given style and a classification model for identifying the style of an utterance.

In recent years, several studies have leveraged the text generation capabilities of rapidly advancing LLMs to address style control. Konen et al. (2024) controlled a style in text generation by adding style vectors to the activation of hidden layers in an LLM. Two types of style vectors were proposed: the training-based and activation-based style vectors. The former trained the style vectors using the cross-entropy loss between the output of the LLM for the empty input token and the target sentence. The latter employed the activation vectors of the layers in the LLM for the given target sentences to obtain the style vector. Li et al. (2024) created a dialog corpus containing utterances in 38 different style categories using an LLM, allowing fine-grained styles to be handled in dialog system development. First, a prompt including the name of a target style is given to the LLM to generate a description of the style and an example sentence. Next, the style description and the example sentence were given to the LLM to generate a rationale that the style of the sentence was consistent with the given style description. Finally, the style name, style description, example sentences, and style rationale as well as a plain context were provided to the LLM to generate a response to the given context in the target style. The constructed dialog corpus consisted of the pairs of the input contexts and the generated responses in different styles.

Although the aforementioned studies can generate natural responses in a specific style, they are limited to considering a single style in style control. In contrast, this study aims to dynamically control multiple styles based on the user's state.

Miura et al. (2024b) proposed a dialog system that flexibly switched between two different styles, the polite style and the casual style, according to the changes in the user's intimacy with the dialog system. The dialog model was trained to generate responses in the polite style when the user's intimacy is low and in the casual style when the intimacy is high, by referring to the intimacy estimation model and two language models of the polite and casual styles. In addition, the style discrimination model was employed to train a dialog model so that the probability of the polite (or casual) style of generated responses, which was estimated by the style discrimination model, became high when the user's intimacy was low (or high). This learning method succeeded in achieving better style control capability than general dialog models. However, there is much room to improve the accuracy of style control due to the poor performance of the intimacy estimation model incorporated in the dialog model. Therefore, this study aims to develop a model for interpreting the user's intimacy by creating user embeddings, so the model could accurately capture the user's intimacy and appropriately perform style control.

3 Proposed Method

3.1 Overview

Figure 1 shows an overview of the dialog model that changes the style based on the recognized user's intimacy. Given a dialog history X, the proposed system generates Y which is a response to X. Here, X is a conversation between a system S and

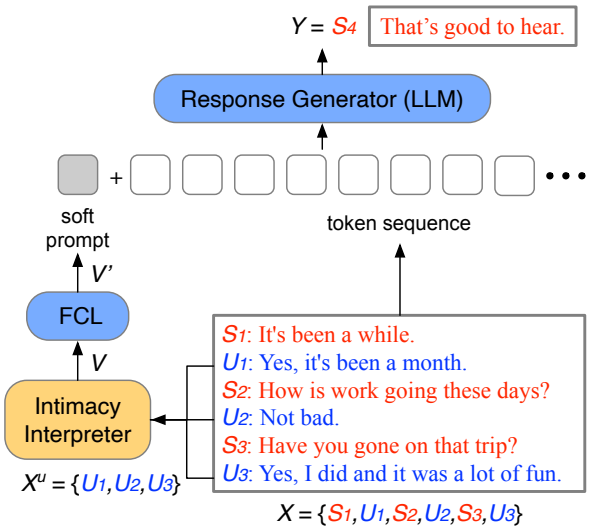

Figure 1: Overview of proposed method

a user U, denoted as $X = \{S_1, U_1, \cdots, S_n, U_n\}$, while Y is the next utterance of the system, i.e., $Y = S_{n+1}$.

The proposed system consists of two submodels. The first is the Intimacy Interpreter. It takes the user's past utterances $X^u = \{U_1, \cdots, U_n\}$ as input and interprets the user's degree of intimacy with the dialog system. The output of the Intimacy Interpreter is an intimacy embedding, a vector representation of the user's intimacy. The second is the Response Generator, which is based on an LLM. It takes the dialog history $X = \{S_1, U_1, \cdots, S_n, U_n\}$ as input and produces a response Y as output. At the beginning of the input token sequence, a soft prompt of the user's intimacy is added. This is a single token embedding derived from the intimacy embedding. Specifically, the size of the intimacy embedding produced by the Intimacy Interpreter is changed to that of the token embeddings of the LLM by the Fully Connected Layer (FCL). It is expected that the response is generated in a casual style when the user's intimacy is high and in a polite style when it is low. The length of the dialog history is 3 in Figure 1, but it can be changed arbitrarily.

The following sections describe the details of the Intimacy Interpreter and Response Generator, respectively.

3.2 Intimacy Interpreter

The Intimacy Interpreter aims to capture the complex and vague nature of the user's intimacy by representing it as an abstract vector. Hereafter, the Intimacy Interpreter is denoted as $P_{II}(V|X^u)$. The model takes as input the n consecutive utterances of a user in a dialog context, $X^u = \{U_1, \cdots, U_n\}$, and outputs a vector V representing the user's intimacy with the dialog system.

This study applies contrastive learning to train the Intimacy Interpreter. An intimacy-labeled dialog corpus D_{in}, where each dialog is labeled with a 5-point Likert scale indicating the level of intimacy of a speaker with a dialog partner, is used for contrastive learning. The details of this corpus are described in 4.1.1. The user's n consecutive utterances in D_{in} are extracted as a sample (X_i^u, IL_i), where IL_i denotes the five-scale intimacy label assigned to the sample X_i^u. Two samples X_i^u and X_j^u are randomly taken from the training data. If the intimacy labels IL_i and IL_j assigned to these two samples are the same, the parameters of the Intimacy Interpreter are updated so that the embedded vectors V_i and V_j become similar. If IL_i and IL_j are not equal, the parameters are updated so that V_i and V_j are different. Specifically, the contrastive loss for training $P_{II}(V|X^u)$ is defined as Equation (1).

$$L_I = \begin{cases} 1 - sim_{cos}(V_i, V_j) & \text{if } IL_i = IL_j \\ |IL_i - IL_j| \cdot \max(0, sim_{cos}(V_i, V_j)) & \text{if } IL_i \neq IL_j \end{cases} \quad (1)$$

$sim_{cos}(\cdot, \cdot)$ represents the cosine similarity between the two sample embedding vectors. When $IL_i \neq IL_j$, the loss becomes large when the difference between IL_i and IL_j is large by giving $|IL_i - IL_j|$ as the weight. The Intimacy Interpreter is obtained by fine-tuning the pre-trained BERT (Devlin et al., 2019) using this loss.

3.3 Response Generator

The Response Generator is denoted as $P_{RG}(Y|V', X)$, where X is the dialog history, V' is the soft prompt derived from the intimacy embedding (V), and Y is the response to be generated. This subsection describes the details of training the Response Generator.

3.3.1 Loss for Style Control

As described earlier, the Response Generator is obtained by fine-tuning an LLM. Following the study of (Miura et al., 2024b), two loss functions, the intimacy-aware word-level loss and the intimacy-aware sentence-level loss, are used to fine-tune the LLM so that the Response Generator generates responses in the appropriate style

(polite or casual) according to the user's intimacy.

Preliminary The intimacy-labeled dialog corpus D_{in} described in subsection 3.2 is also used to train the the Response Generator. In addition, two style corpora are prepared to handle polite and casual styles in response generation. One is C_{po} which consists of polite style sentences, and the other is C_{ca} which consists of the casual sentences.

Before the training of the Response Generator, an intimacy estimation model $P(I|X^u)$ is trained in advance. This model predicts I, the user's level of intimacy with a dialog system, given the user's past n utterances (X^u) as input. In our model, I is defined as either "low" or "high". The intimacy estimation model is pre-trained using D_{in}. Note that this is a different model from the Intimacy Interpreter $P_{II}(V|X^u)$. The Intimacy Interpreter produces the intimacy embedding, while the intimacy estimation model is a binary classifier.

Intimacy-aware Word-Level Loss Two style language models are pre-trained. A polite style language model $P_{po}(T)$ is trained using C_{po}, and a casual style language model $P_{ca}(T)$ is trained using C_{ca}. These models evaluate how likely the given sentence T is in the polite or casual style. They are employed to calculate the polite style word-level loss L_w^{po} and the casual style word-level loss L_w^{ca}, respectively, as shown in Equation (2).

$$L_w^s = d(\mathbf{p_Y}||\hat{\mathbf{p}}_\mathbf{Y}) \stackrel{\text{def}}{=} \sum_{i=1}^{m} D_{KL}(p_{y_i}||\hat{p}_{y_i}), \quad (2)$$

where s denotes the style, either po (polite) or ca (casual). This loss is computed for each dialog sample (X, Y) in the training data. Y is denoted as a token sequence $\{y_1, \cdots, y_m\}$. Let $\mathbf{p_Y} = \{p_{y_1}, \cdots, p_{y_m}\}$ be the distribution of the predicted probability of the next word given by the dialog model $P_{RG}(Y|V', X)$, and $\hat{\mathbf{p}}_\mathbf{Y} = \{\hat{p}_{y_1}, \cdots, \hat{p}_{y_m}\}$ be the probability distribution predicted by the style language model $P_s(T)$. D_{KL} is the Kullback-Leibler divergence of the two probability distributions, indicating whether the words generated by the dialog model follow the specified (polite or casual) style.

As shown in Equation (3), the intimacy-aware word-level loss is defined as the weighted sum of two losses, where $p(I{=}\text{low}|X^u)$ is the weight for L_w^{po} and $p(I{=}\text{high}|X^u)$ is the weight for L_w^{ca}. $p(I{=}\text{low}|X^u)$ and $p(I{=}\text{high}|X^u)$ are the probabilities of the low intimacy and high intimacy classes, respectively, predicted by the intimacy estimation model.

$$L_w^{in} \stackrel{def}{=} p(I{=}\text{low}|X^u) \cdot L_w^{po} + p(I{=}\text{high}|X^u) \cdot L_w^{ca} \quad (3)$$

It is expected that this loss will cause the Response Generator to generate more polite style tokens when the intimacy is low, and more casual style tokens when the intimacy is high.

Intimacy-aware Sentence-Level Loss First, we train a style discrimination model $P'(S|T)$ that classifies the style S of a sentence T. The style S is either polite or casual. The style discrimination model is pre-trained from training data in which utterances in C_{po} are samples of the polite class and utterances in C_{ca} are samples of the casual class.

Let $\hat{Y}$ be the response generated by $P_{RG}(Y|V', X)$. The style of $\hat{Y}$ is identified using the style discrimination model $P'(S|T)$, and the $p(S{=}\text{polite}|\hat{Y})$ and $p(S{=}\text{casual}|\hat{Y})$, the predicted probabilities of the polite and casual classes respectively, are calculated. The intimacy-aware sentence-level loss L_s^{in} is defined as the weighted sum of the logarithms of these probabilities, as shown in Equation (4).

$$\begin{aligned} L_s^{in} \stackrel{def}{=} & -p(I{=}\text{low}|X^u) \cdot \log p(S{=}\text{polite}|\hat{Y}) \\ & -p(I{=}\text{high}|X^u) \cdot \log p(S{=}\text{casual}|\hat{Y}) \end{aligned} \quad (4)$$

This loss will contribute to making the Response Generator to generate polite (or casual) style sentences when intimacy is low (or high).

3.3.2 Negative Log-likelihood Loss

The two losses described in 3.3.1 are designed to maintain style consistency. A model fine-tuned solely by these losses may exhibit inconsistency between the dialog context and the generated response. Therefore, a common loss for training dialog models, the negative log-likelihood loss defined as shown in Equation (5), is also used. The value $p(Y|V', X)$ denotes the probability of the ground-truth response Y in the training data being generated by the Response Generator for a given soft prompt of user's intimacy V' and the dialog context X.

$$L_{NLL} = -\log\ p(Y|V', X) \quad (5)$$

3.3.3 Training Objective

The loss for training the Response Generator, L_D, is a weighted sum of two losses for style control (L_w^{in} and L_s^{in}) and a loss for content generation (L_{NLL}) as follows:

$$L_D = \beta_w \cdot L_w^{in} + \beta_s \cdot L_s^{in} + \beta_{NLL} \cdot L_{NLL} \quad (6)$$

The weights β_w, β_s, and β_{NLL} are hyperparameters.

3.4 Training Details

Our entire dialog model, shown in Figure 1, is trained based on two losses: L_I and L_D. On the one hand, the parameters of the Intimacy Interpreter $P_{II}(V|X^u)$ are updated using L_I. On the other hand, the parameters of the Response Generator $P_{RG}(Y|V', X)$ and the FCL that transforms the dimension of the intimacy embedding are updated using L_D.[1] The Response Generator is based on the LLM, which is computationally expensive to fine-tune. Therefore, LoRA (Hu et al., 2022) is applied to fine-tune the Response Generator.

4 Experiments

4.1 Datasets

4.1.1 Dialog Corpus with Intimacy Label

The JID corpus (Miura et al., 2024a) is used as the intimacy-labeled corpus D_{in}. This corpus consists of recorded and transcribed conversations of about 10 minutes between two speakers. For each conversation, the intimacy labels of each of the two speakers are annotated using a five-point Likert scale. The number of subjects who participated in the dialogs is 19, the number of dialogs is 54, and the total number of utterances is 6,984.

The 54 dialogs in the JID corpus are divided into three subsets: a training set of 33 dialogs, a validation set of 9, and a test set of 12. As mentioned in section 3, the dialog model accepts the preceding dialog context of the user and the system, $X = \{S_1, U_1, \cdots, S_n, U_n\}$, as input and generates the subsequent response S_{n+1} as output. Hereafter, the pair of a dialog context and its corresponding response, denoted by (X, S_{n+1}), will be referred to as "response instance." The first $n \times 2$ utterances and the next utterance in a dialog are extracted as (X, S_{n+1}). One speaker in the corpus is designated as the system and the other as the user. This procedure is then repeated with the utterance shifted one by one to obtain multiple response instances. In this experiment, the context length is set to $n = 3$. The statistics of the dataset are shown in Table 1.

4.1.2 Style Corpus

Two style corpora of the polite and casual style, C_{po} and C_{ca}, are required to train style language models and a style discrimination model. The KeiCO corpus (Liu and Kobayashi, 2022) is used as C_{po}. This corpus contains utterances using various types of honorific expressions in Japanese. Besides, C_{ca} is constructed by extracting utterances from conversations between speakers who know each other in the BTSJ Japanese Natural Conversation corpus (Usami, 2021). C_{po} and C_{ca} contain 7,324 and 13,521 utterances, respectively.

[1]The blue modules in Figure 1 indicate the models trained with the loss L_D.

	Training	Validation	Test
Dialog	33	9	12
Response Instance	4,032	921	1,284

Table 1: Statistics of Dataset

次の対話文脈に対して、あなたはBとして応答を生成してください。
(For the following dialog context, generate a response as B.)

【Dialog Context】

Figure 2: Template of Zero-shot Prompt

次の対話文脈に対して、あなたはBとして応答を生成してください。
ただし、AがBに抱く親密度を推測して、親密度が低い場合は丁寧なスタイルで、親密度が高い場合はカジュアルなスタイルで応答を生成してください。
(For the following dialog context, generate a response as B. Guess the level of intimacy A has with B and generate a response in a polite style if the level of intimacy is low and in a casual style if the level of intimacy is high.)

【Dialog Context】

Figure 3: Template of Style Control Prompt

---------1st step
この対話からAがBに抱く親密度は
(From this dialog, the level of intimacy that A feels towards B is)

【Dialog Context】

---------2nd step
次の対話文脈に対して、あなたはBとして応答を生成してください。
ただし、「【output of the first step】」という解釈を踏まえて、親密度が低い場合は丁寧なスタイルで、親密度が高い場合はカジュアルなスタイルで応答を生成してください。
(With the interpretation of 【output of the first step】, generate responses in a polite style if the level of intimacy is low, and in a casual style if the level of intimacy is high.)

【Dialog Context】

Figure 4: Template of Two-step Prompt

4.2 Experimental Settings

The following methods are compared in the experiment.

- **Zero-shot prompt (Zero-shot)** This method uses an LLM as a dialog model without fine-tuning or prompting for style control. We only give an instruction for generating responses to the input dialog context. The details of the prompt are shown in Figure 2.
- **Zero-shot prompt for style control (Style**

control prompt) This method uses a pre-trained LLM as a dialog model, where a prompt is given to instruct the LLM to generate utterances taking the style control into account. The details of the prompt are shown in Figure 3.

- **Two-step prompt (Two-step)** This method uses a pre-trained LLM as a dialog model using two sequential prompts. We first instruct the LLM to infer the user's level of intimacy, and then to generate the system's response in a polite or casual style according to the inferred level of intimacy. See Figure 4 for details.

- **STYLEDGPT** This is a model where the style is controlled by STYLEDGPT (Yang et al., 2020). Specifically, we fine-tune the LLM to generate utterances that are consistent with the style of the entire JID corpus. The style language model is trained on training data from the JID corpus. The style discrimination model, which distinguishes whether an utterance is in the style of the JID corpus, is trained using utterances from the JID corpus as positive samples and sentences from Japanese Wikipedia as negative samples.

- **Ours$_{\text{auto}}$** This is our proposed method described in section 3.

- **Ours$_{\text{gold}}$** Our proposed method where the gold intimacy labels in the JID corpus are used instead of the prediction by the intimacy estimation model. When calculating the losses in Equation (3) and (4), $p(I{=}\text{low}|X^u)$ and $p(I{=}\text{high}|X^u)$ given as follows.

$$p(I{=}\text{low}|X^u) = 1 - \frac{IL}{5} \quad (7)$$
$$p(I{=}\text{high}|X^u) = \frac{IL}{5} \quad (8)$$

IL represents the five-level intimacy label assigned to X^u in the JID corpus. This model evaluates our approach of considering the user's intimacy for the appropriate style control under the ideal condition where the user's intimacy is correctly predicted.

4.3 Implementation Details

4.3.1 Intimacy Interpreter and Response Generator

The Intimacy Interpreter described in subsection 3.2 is obtained by contrastive learning based on the Japanese BERT model[2], which was pre-trained on large-scale corpora of Japanese Wikipedia and Japanese CC-100.

The Response Generator described in subsection 3.3 is obtained by fine-tuning llm-3-3.7b[3], which is an LLM based on Transformer (Vaswani et al., 2017) and has been trained on various large Japanese datasets. We also adopted llm-3-3.7b as the LLM for other baseline dialog models.

For the hyperparameters during training, the learning rate for the Intimacy Interpreter is $1e^{-6}$, while that for the Response Generator is $1e^{-20}$. For both models, the batch size is 4 and the number of epochs is 5. These values were optimized on the validation set according to the StyCor criteria, which will be defined in subsection 4.5. The Adam optimizer was used to learn the models.

The hyperparameters β_w, β_s, and β_{NLL} in Equation (6) are set to 0.5, 1, and 0.005, respectively. These values are determined so that the influence of the three types of losses is uniform. Specifically, we calculate the average of the absolute value of each of the three losses in the training data and then determine the weight of each loss as the approximate inverse ratio of the average to the minimum value.

4.4 Other Submodels

Several submodels are pre-trained before training of the Intimacy Interpreter and Response Generator.

The style language models $P_{po}(T)$ and $P_{ca}(T)$ are obtained by fine-tuning GPT-2. We use the pre-trained model japanese-gpt2-medium[4], which has been trained on a large Japanese dialog dataset. All utterances in C_{po} and C_{ca} are used to train $P_{po}(T)$ and $P_{ca}(T)$, respectively. The learning rate is set to $5e^{-4}$, the batch size to 4, and the epoch to 20. The Adam optimizer is used to fine-tune the models.

The style discrimination model $P'(S|T)$ is obtained by fine-tuning the Japanese BERT model[2]. A total of 20,575 utterances are used, comprising 7,274 polite utterances in C_{po} and 13,301 casual utterances in C_{ca}. The learning rate is set to $1e^{-7}$, the batch size to 128, and the epoch to 10. The Adam optimizer is used to fine-tune the model. The accuracy of the style discrimination model was 99% when it was evaluated on the 100 test utterances

[2]https://huggingface.co/tohoku-nlp/bert-large-japanese-v2

[3]https://huggingface.co/llm-jp/llm-jp-3-3.7b

[4]https://huggingface.co/rinna/japanese-gpt2-medium

(50 polite and 50 casual) that were not used for training.

The intimacy estimation model $P(I|X^u)$ is based on the Japanese BERT model[2]. The JID corpus is used for fine-tuning the BERT. The learning rate is set to $5e^{-6}$, the batch size to 1, and the epoch to 10. The Adam optimizer is used to train the model. The accuracy of the intimacy estimation model on the test data was 69%.

4.5 Evaluation Criteria

Both automatic and human evaluations are carried out to assess responses generated by various methods.

4.5.1 Automatic Evaluation

In the automatic evaluation, the quality of the generated responses is evaluated from three perspectives: relevance, diversity, and style. The relevance is measured by BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). Specifically, the similarity between a generated response and a ground-truth response is evaluated using BLEU-1, BLEU-2, ROUGE-1, and ROUGE-2. The diversity is measured by Distinct-1 (Dist-1) and Distinct-2 (Dist-2), following the experiment of (Li et al., 2016). The style is evaluated by measuring "Style Correlation" (StyCor). The StyCor metric is defined as the correlation between the probability of the casual style $p(S$=casual$|Y)$ and the ground-truth level of the intimacy.[5] This correlation is high when both the predicted probability of the casual style and the level of intimacy are high, or both are low (i.e., the probability of the polite style is high and the level of intimacy is low).

4.5.2 Human Evaluation

The quality of the generated responses is evaluated by humans. To reduce the burden on evaluators, STYLEDGPT and Ours$_{\text{auto}}$ are excluded from the human evaluation. A hundred response instances are randomly taken from the test set of the JID corpus. The dialog context X of each response instance is used as input, and a response is generated using the dialog models. Subjects evaluate these responses from the following three perspectives.

- Style Control: Does the response align with the appropriate style for the relationship between the two speakers? Annotators are also instructed to read the dialog context and guess the relationship between the speakers.
- Relevance: Is the content of the response relevant and consistent with the context?
- Fluency: Is the response natural, fluent, and free of grammatical errors?

For each item, the quality of the responses was assessed by giving a score on a 5-point Likert scale from 1 (inappropriate) to 5 (appropriate). Five native Japanese speakers participated in the manual evaluation. Agreement between annotators' scores was measured using Fleiss's kappa (Fleiss and Jacob, 1973).

5 Results

5.1 Results of Automatic Evaluation

The results of the automatic evaluation are shown in Table 2. The StyCor of Ours$_{\text{auto}}$ and Ours$_{\text{gold}}$ were 0.239 and 0.250, respectively, outperforming other baseline methods. This confirms that the proposed method, which adjusts the style based on the level of intimacy, can effectively control the polite and casual styles. The decrease of StyCor of Ours$_{\text{auto}}$ compared to Ours$_{\text{gold}}$ may be due to the low accuracy of the intimacy estimation model.

In the evaluation of the relevance, STYLEDGPT and our proposed models achieved better BLEU and ROUGE scores than other baselines, since these models are fine-tuned using the JID corpus, which was the same domain as the test data. However, our models performed slightly worse than STYLEDGPT. On the other hand, the diversity (Dist-1 and Dist-2) of all models was high.

Although the BLEU and ROUGE of our method are worse than those of STYLEDGPT, we think that these indicators are only for reference in automatic evaluation. BLEU and ROUGE only evaluate the similarity between the generated and ground-truth responses, while there could be other appropriate responses that are not included in the dataset. On the other hand, our proposed method clearly outperforms STYLEDGPT in terms of StyCor, indicating superior style control capabilities.

To sum up, our models can improve the ability of the style control with a little decrease in relevance.

5.2 Results of Human Evaluation

The results of the human evaluation are shown in Table 3. The "Score" column shows the average

[5]The five-scale score is normalized to values between 0 and 1.

Method	Relevance				Diversity		Style
	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	Dist-1	Dist-2	StyCor
Zero-shot	0.0483	0.0034	0.0780	0.0044	0.942	0.978	0.164
Style control prompt	0.0578	0.0053	0.1014	0.0073	**0.965**	**0.991**	0.207
Two-step	0.0575	0.0028	0.0932	0.0041	0.946	0.984	0.162
STYLEDGPT	0.2520	**0.1571**	**0.3392**	**0.2108**	0.925	0.935	0.171
$\text{Ours}_{\text{auto}}$	0.2067	0.1205	0.2986	0.1725	0.895	0.900	0.239
$\text{Ours}_{\text{gold}}$	**0.2544**	0.1463	0.3390	0.1999	0.925	0.930	**0.250**

Table 2: Results of Automatic Evaluation

Method	Style Control			Relevance			Fluency		
	Score	κ	p	Score	κ	p	Score	κ	p
Zero-shot	4.31	0.50	$1e^{-6}*$	4.13	0.50	0.086	4.51	0.68	$9e^{-11}*$
Style control prompt	4.34	0.51	$9e^{-5}*$	4.22	0.53	0.553	4.63	0.72	$8e^{-7}*$
Two-step	4.33	0.51	$6e^{-5}*$	4.02	0.44	0.002*	4.49	0.69	$2e^{-12}*$
$\text{Ours}_{\text{gold}}$	4.61	0.60	–	4.26	0.54	–	4.86	0.84	–

Table 3: Results of Human Evaluation. * means $p < 0.05$.

score of the five subjects, while the "κ" column indicates Fleiss's κ. Welch's test is performed to verify whether there was a significant difference in the scores between $\text{Ours}_{\text{gold}}$ and other methods. The "p" column represents the p-value of this statistical test.

For Style Control, $\text{Ours}_{\text{gold}}$ received the highest score. Additionally, significant differences with all other methods were confirmed. This demonstrates the effectiveness of the approach proposed in this study, which considers the user's level of intimacy for the appropriate selection of polite and casual styles. The κ value was 0.60, which indicated moderate agreement.

In terms of Relevance, $\text{Ours}_{\text{gold}}$ achieved the highest score. However, significant differences were only observed when compared to Two-step. The proposed method performed comparably to the baseline methods in generating responses relevant to the dialog context.

The Fluency score of the proposed method was significantly higher than the other models, indicating its superior ability to generate natural utterances.

6 Ablation Study

Table 4 shows the results of the ablation study. The Ours-SCL is the model where two intimacy-aware style control losses, L_w^{in} and L_s^{in}, are removed from Equation (6). The Ours-II indicates the removal of the Intimacy Interpreter, which is almost equivalent to the dialog model presented in (Miura et al., 2024b).[6] This model is trained using the gold intimacy labels to calculate the loss L_D, so the above two models are compared to $\text{Ours}_{\text{gold}}$.

The results demonstrated that both the use of the style control losses and the incorporation of the Intimacy Interpreter could improve the StyCor score. Especially, a significant decrease was found in Ours-SCL, indicating that the intimacy-aware style control losses are effective in changing the style appropriately. On the other hand, the contribution of the Intimacy Interpreter was rather limited. It should be noted that both the style control losses and the Intimacy Interpreter could also improve the relevance and diversity of the generated responses.

7 Conclusion

In this paper, we proposed the novel method to control the style of a dialog system based on the user's level of intimacy. The model that interpreted the user's level of intimacy was incorporated into the dialog model. This Intimacy Interpreter was trained by contrastive learning using the dialog corpus annotated with the intimacy labels. Furthermore, based on the LLM, which had an excellent capability to generate general responses, we applied two loss functions to improve the model's ability to control the style. The results of both au-

[6]The base LLMs are different: llm-jp-3-3.7b was used in this paper, while GPT-2 was used in (Miura et al., 2024b).

Methods	Relevance				Diversity		Style
	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	Dist-1	Dist-2	StyCor
Ours-SCL (w/o style control loss)	0.2105	0.1175	0.2954	0.1697	0.879	0.889	0.200
Ours-II (w/o intimacy interpreter)	0.2170	0.1257	0.3086	0.1826	0.907	0.917	0.247
$\text{Ours}_{\text{gold}}$	**0.2544**	**0.1463**	**0.3390**	**0.1999**	**0.925**	**0.930**	**0.250**

Table 4: Results of Ablation Study

tomatic and human evaluations demonstrated that the proposed method outperformed the baseline in generating responses in a casual style when the user's level of intimacy was high and in a polite style when it was low.

The proposed dialog model was trained using a dialog corpus annotated with the speaker's level of intimacy. However, the availability of such a corpus is rather limited, while the construction of new corpora requires considerable costs. Therefore, it is essential to explore ways to enable LLMs to acquire the ability to control the style without relying on the intimacy-labeled corpus. Another important future work is to explore new style control frameworks that do not rely on pre-training the style language models and/or the style discrimination model.

References

Noora Aapakallio. 2021. *Understanding Through Politeness – Translations of Japanese Honorific Speech to Finnish and English*. University of Eastern Finland.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Joseph L. Fleiss and Cohen Jacob. 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring Latent Spaces for Stylized Response Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1814–1823, Hong Kong, China. Association for Computational Linguistics.

Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Reina Akama. 2021. *Dialogue System Live Competition: Identifying Problems with Dialogue Systems Through Live Event*, pages 185–199. Springer Singapore.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yukiko Kageyama, Yuya Chiba, Takashi Nose, and Akinori Ito. 2018. Improving User Impression in Spoken Dialog System with Gradual Speech Form Control. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 235–240, Melbourne, Australia. Association for Computational Linguistics.

Chandra Khatri, Anu Venkatesh, Behnam Hedayatnia, Raefer Gabriel, Ashwin Ram, and Rohit Prasad. 2018. Alexa Prize — State of the Art in Conversational AI. *AI Magazine*, 39(3):40–55.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style Vectors for Steering Generative Large Language Models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802, St. Julian's, Malta. Association for Computational Linguistics.

Jinpeng Li, Zekai Zhang, Quan Tu, Xin Cheng, Dongyan Zhao, and Rui Yan. 2024. Stylechat: Learning recitation-augmented memory in llms for stylized dialogue generation. *arXiv preprint arXiv:2403.11439*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Muxuan Liu and Ichiro Kobayashi. 2022. Construction and Validation of a Japanese Honorific Corpus Based on Systemic Functional Linguistics. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 19–26, Marseille, France. European Language Resources Association.

Takuto Miura, Kiyoaki Shirai, Hideaki Kanai, and Natthawut Kertkeidkachorn. 2024a. Construction of a Japanese Dialog Corpus Annotated with Speakers' Intimacy. In *The 38th Pacific Asia Conference on Language, Information and Computation (PACLIC 38)*, page 10, Tokyo, Japan. Association for Computational Linguistics.

Takuto Miura, Kiyoaki Shirai, and Natthawut Kertkeidkachorn. 2024b. Intimacy-aware Style Control in Dialog Response Generation. In *The 9th Linguistic and Cognitive Approaches to Dialog Agents Workshop (LACATODA 2024)*, pages 5–16, Kyoto, Japan. CEUR-WS.

Tong Niu and Mohit Bansal. 2018. Polite Dialogue Generation Without Parallel Data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational AI: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Michael Silverstein. 2003. Indexical Order and the Dialectics of Social Life. *Language & Communication*, 23:193–229.

Mayumi Usami, editor. 2021. *BTSJ-Japanese Natural Conversation Corpus with Transcripts and Recordings (March 2021)*. National Institute for Japanese Language and Linguistics, Japan.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ronald Wardhaugh and Janet M Fuller. 2021. *An introduction to sociolinguistics*. John Wiley & Sons.

Ze Yang, Wei Wu, Can Xu, Xinnian Liang, Jiaqi Bai, Liran Wang, Wei Wang, and Zhoujun Li. 2020. StyleDGPT: Stylized Response Generation with Pretrained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1548–1559, Online. Association for Computational Linguistics.

Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2021. Stylized dialogue response generation using stylized unpaired texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14558–14567, Online. AAAI Press.

Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. 2021. Neural stylistic response generation with disentangled latent variables. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4391–4401, Bangkok, Thailand. Association for Computational Linguistics".