Intimacy-aware Style Control in Dialog Response Generation

Takuto Miura*, Kiyoaki Shirai and Natthawut Kertkeidkachorn

Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, 9231211, Japan

Abstract

One of the crucial features in developing a dialog system is the choice of an appropriate speech style. This paper proposes a novel method for training a dialog model that can effectively control the style of a response. 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 shows a high level of intimacy. Using a pre-trained language model (PLM) as a base dialog model, two loss functions are proposed for fine-tuning the PLM to generate responses in an appropriate style. One is the intimacy-aware word-level loss, which serves to ensure that the dialog model generates a polite or casual word when the user's level of intimacy is low or high. The other is the intimacy-aware sentence-level loss, which functions to increase the probability of the polite style of the generated utterance when the user's level of intimacy is low, and vice versa. The results of both automatic and human evaluations in the experiments demonstrate that the proposed method is more effective than the baselines in generating responses that align with the user's degree of intimacy. Furthermore, the proposed method exhibits comparable relevance and fluency to the PLM, indicating that the losses for the style control do not diminish the PLM's exceptional capacity for generating relevant and fluent responses.

Keywords

Dialog System, Speech Style, Intimacy

1. Introduction

Dialog systems that freely chat with users on a wide range of topics have attracted a great deal of attention in recent years [1, 2, 3]. These systems are required to have comfortable conversations with users and build long-term friendly relationships with them [4]. Humans adjust their speech style according to their social relationships with their partners and/or the level of intimacy they share with their partners [5, 6, 7]. Such behavior is referred to as a "style control" hereafter. One of the style controls is to use both polite and casual styles depending on the relationship with the partner [8, 9]. Polite styles are often used in a conversation with a boss or a teacher, while casual styles are often employed with a friend or a life partner. The style control should be considered in all conversations, whether between humans or between humans and dialog systems [10].

The goal of this research is to develop a dialog system that flexibly controls speech styles according to the user. Specifically, concerning the user's intimacy with the dialog system, a response is generated in a polite style when the user's level of the intimacy is low, and in a casual style when the level of the intimacy is high. To achieve this, we propose a method to incorporate knowledge necessary for style control by fine-tuning a dialog model based on a pre-trained language model (PLM) that is capable of generating a variety of responses consistent with the dialog context. A new loss function for fine-tuning a dialog model is designed so that the model generates polite or casual responses when the level of the intimacy is low or high, where the level of the intimacy is estimated from the user's past utterances.

The contributions of this paper are summarized as follows:

- We develop a dialog system that estimates the user's level of the intimacy and controls the polite and casual styles in generating responses accordingly.

The 9th Linguistic and Cognitive Approaches to Dialog Agents Workshop

*Corresponding author.

s2460005@jaist.ac.jp (T. Miura); kshirai@jaist.ac.jp (K. Shirai); natt@jaist.ac.jp (N. Kertkeidkachorn)
0009-0008-1178-7407 (T. Miura)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3862/paper1.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

- We propose an approach to incorporate knowledge for style control into an existing outstanding PLM-based dialog model.
- We demonstrate the effectiveness of the proposed method by both automatic and manual evaluations.

2. Related Work

Several methods have been developed for the generation of responses in a specific style. Niu and Bansal defined the task of generating responses in a predefined style, such as polite or rude [11]. Gao et al. proposed a method that shared the latent space between conversational and stylistic modeling and developed a model that generated responses in a specified style while maintaining consistency with the dialog context [12]. Zhu et al. extended Gao's model so that the representation of content and style was learned in different dimensions in latent space [13]. Zheng et al. proposed a method for automatic construction of a dialog corpus consisting of utterances in a certain style, aiming to train a stylized dialog model [14]. Specifically, they created a Seq2Seq model, which transforms sentences in an original dialog corpus into ones in the specified style, using texts written in that style. Tsai et al. evaluated three approaches to achieve both content and style fidelity: conditional learning, guided fine-tuning, and guided decoding [15]. In conditional learning, special tokens about a style are added to the input of the dialog model. In guided fine-tuning, a style of an utterance is classified, and the classification result is added to the input of the dialog model. In guided decoding, the weights of the output of the decoder are determined based on the result of the style classification model. Saha et al. proposed a multitask learning method that predicts the speaker's personality and intention when training a dialog model [16]. This approach is designed to control the style following the predicted state of the speaker.

Based on the aforementioned studies on maintaining a style in response generation, more recent methods have been developed to add the capability of style control to a well-developed existing dialog model. Sun et al. trained a dialog model using reinforcement learning, in which responses similar to the ground-truth response and including style-related tokens got a higher reward [17]. The similarity between responses was measured by the cosine similarity of the sentence embeddings, while the style-specific tokes were identified by the pre-trained classification model. Li et al. retrieved a sentence similar to an utterance from a corpus of sentences written in a specific style and fed the retrieved sentence and the utterance into a dialog model to generate a stylized response [18]. Since the retrieved style sentence might be harmful to generate a response consistent with the context, they incorporated an encoder that removed features not pertinent to the context, resulting in the extraction of style features only, into the dialog model. This encoder was trained simultaneously with the dialog model. Yang et al. proposed loss functions using a language model that generated sentences in the specified style and a classification model that identified the style of a sentence for fine-tuning the PLM of the dialog model [19].

Although the above previous studies can generate natural stylized responses, they are limited in their ability to handle a single style. In contrast, our method enables the control of multiple styles according to the user's mental state.

Several studies focused on the emotional state of the user during a dialog. Skowron et al. showed that interactive expression of emotions in response to the user's feelings can significantly contribute to enhancing the enjoyment of the chat and the emotional connection between the user and the system [20]. D'mello and Graesser developed an intelligent tutoring agent that responds empathetically or motivationally according to the user's cognitive and emotional states [21]. This interactive agent dramatically improved the learning efficacy of students with limited domain knowledge. Thus, controlling the type of response of the system according to the user's internal state exerts a considerable influence. In this study, we deal with the user's intimacy as the user's internal state and the speech style as the type of response.

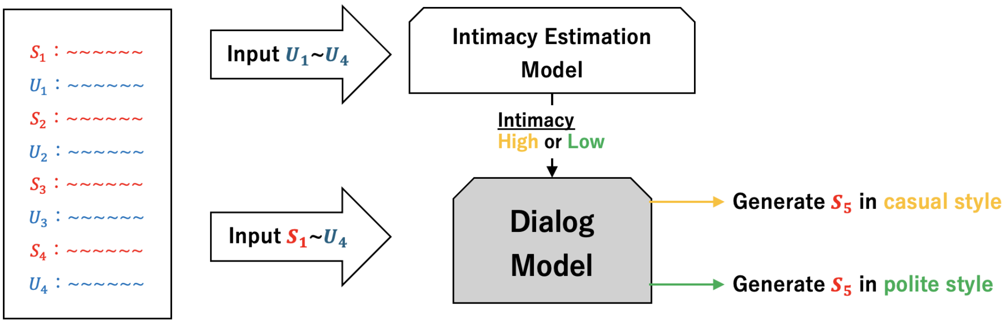

Figure 1: Overview of proposed method

3. Proposed Method

The proposed method learns a dialog model that can adapt its style to either polite or casual by the user's level of intimacy, which is automatically estimated from the user's historical utterances. Figure 1 shows an overview of the proposed method. Let us suppose that a dialog model generates a response to a user for a given dialog context $X = \{S_1, U_1, \cdots, S_n, U_n\}$, where a system ($S$) and a user ($U$) make an utterance alternately. Figure 1 exemplifies the case of n = 4. The intimacy estimation model employs the user's previous utterances $X_u = \{U_1, \cdots, U_n\}$ as input and determines whether the user's level of intimacy with the dialog system is high or low. The dialog model accepts the context X and the estimated intimacy level as input and generates the response S_{n+1} in a casual style when the user's intimacy is high and in a polite style when the user's intimacy is low.

To learn the above dialog model, we extend STYLEDGPT [19], a model that consistently generates responses in a specified style obtained by fine-tuning a PLM that can generate versatile responses. To avoid impairing the exceptional response generation capability of the PLM, only the loss function in fine-tuning is modified while the architecture of the PLM remains intact. Indeed, Yang et al. demonstrated that STYLEDGPT performed well not only in its ability to produce utterances in the specified style but also generate relevant and fluency responses [19]. First, we provide an overview of STYLEDGPT in subsection 3.1 and then describe the details of the proposed method in the succeeding subsections.

3.1. STYLEDGPT

STYLEDGPT employs DialoGPT [22] as a PLM and learns a model that consistently generates responses in a specified style by fine-tuning it. DialoGPT is a Seq2Seq model based on GPT-2 [23] and has been pre-trained with a large amount of dialog data.

Word-Level Loss First, a style language model $P_s(T)$ is trained in advance. Using a style corpus, D_{style}, consisting of only texts in a given style, GPT-2 is trained as an autoencoder, i.e., the same sentence $T \in D_{style}$ is given as input and output for fine-tuning.

Let $P(Y|X)$ be a dialog model that returns a response Y for a given dialog context X. The loss is computed for each dialog sample (X, Y) in the training data D_{dialog}. Y is a sequence of words denoted by $Y = \{y_1, \cdots, y_m\}$. 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 $P(Y|X)$. Also, let $\hat{\mathbf{p}}_\mathbf{Y} = \{\hat{p}_{y_1}, \cdots, \hat{p}_{y_m}\}$ be the distribution of the probability of predicting the next word given by the style language model $P_s(Y)$ when the output Y of the dialog model is taken as input of the style language model. The distance between $\mathbf{p_Y}$ and $\hat{\mathbf{p}}_\mathbf{Y}$ is defined as word-level loss L_w as in Eq. (1).

$$L_w = d(\mathbf{p_Y} || \hat{\mathbf{p}}_\mathbf{Y}) \overset{\text{def}}{=} \sum_{i=1}^{m} D_{KL}(p_{y_i} || \hat{p}_{y_i}) \tag{1}$$

D_{KL} is the Kullback-Leibler (KL) divergence of the two probability distributions. This loss causes $\mathbf{p_Y}$ to approach $\hat{\mathbf{p}}_\mathbf{Y}$, i.e., the dialog model is trained to produce utterances in the specified style.

Sentence-Level Loss First, a style discrimination model $P(S|T)$ is trained in advance. It identifies whether a sentence T is written in a given style S. This model is trained on a dataset that consists of D_{style}, a corpus of sequences written in the specific style, as positive samples, and D_{dialog}, a general dialog corpus, as negative samples.

The loss is computed for each dialog sample $(X, Y) \in D_{dialog}$. Let $\hat{Y}$ be a response generated by the dialog model $P(Y|X)$ for the input X, and $p(S|\hat{Y})$ be the probability that the style of $\hat{Y}$ is coincident with the style S. Then, the sentence-level loss L_s is defined as in Eq. (2).

$$L_s = -\log\ p(S|\hat{Y}) \tag{2}$$

This loss causes the dialog model $P(Y|X)$ to produce utterances in the style S.

Negative Log-likelihood Loss The two losses mentioned above are designed to take into account the style of a response. Fine-tuning a model with only these losses may result in a lack of consistency between a context and a generated response. Therefore, the negative log-likelihood loss (Eq. (3)) is also used, which is a common loss for training a dialog model. $p(Y|X)$ is the probability that the dialog model generates a ground-truth response Y from X, where (X, Y) is a sample in D_{dialog}.

$$L_{NLL} = -log\, p(Y|X) \tag{3}$$

3.2. Loss for Style Control

We modify the word-level and sentence-level losses in STYLEDGPT to control the style of the response according to the user's level of intimacy.

First, an intimacy estimation model $P(I|X_u)$ is trained. 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 a dialog corpus annotated with the speaker's intimacy.

To handle both polite and casual styles in response generation, two style corpora are prepared. One is D^{po}_{style}, which consists of polite-style sentences, and the other is D^{ca}_{style}, which consists of casual-style sentences.

Intimacy-aware Word-Level Loss First, the language models of the polite and casual styles, $P_{po}(T)$ and $P_{ca}(T)$, are pre-trained using the corpora D^{po}_{style} and D^{ca}_{style}, respectively. Next, the word-level loss of the polite style, L^{po}_w, is computed as in Eq. (1). It evaluates how likely a response is to be polite. Similarly, the word-level loss of the casual style, L^{ca}_w, is calculated. Finally, the intimacy-aware word-level loss, L^{in}_w, is defined as the weighted sum of these two losses (Eq. (4)). $p(I\text{=low}|X_u)$ and $p(I\text{=high}|X_u)$ are the weights for L^{po}_w and L^{ca}_w, which are the probability that the user's level of intimacy is low and high, respectively.

$$L^{in}_w \stackrel{def}{=} p(I\text{=low}|X_u) \cdot L^{po}_w + p(I\text{=high}|X_u) \cdot L^{ca}_w \tag{4}$$

This loss is expected to encourage the generation of more polite tokens when intimacy is low and more casual tokens when intimacy is high.

Intimacy-aware Sentence-Level Loss First, we train a style discrimination model $P'(S|T)$ that identifies a style S of a sentence T, where S is either polite or casual. The style discrimination model is trained in advance on training data where utterances in D^{po}_{style} are samples of the polite class and those in D^{ca}_{style} are samples of the casual class.

Let $\hat{Y}$ be the output of the dialog model $P(Y|X)$ for a given context X. Then, the style of $\hat{Y}$ is identified by the style discrimination model and $p(S\text{=polite}|\hat{Y})$ and $p(S\text{=casual}|\hat{Y})$ are obtained. Following the sentence-level loss of STYLEDGPT, the intimacy-aware sentence-level loss, L^{in}_s, is defined as the weighted sum of the logarithms of these probabilities, using the two probabilities $p(I\text{=low}|X_u)$ and $p(I\text{=high}|X_u)$ as weights (Eq. (5)).

$$L^{in}_s \stackrel{def}{=} -p(I\text{=low}|X_u) \cdot \log p(S\text{=polite}|Y) - p(I\text{=high}|X_u) \cdot \log p(S\text{=casual}|Y) \tag{5}$$

This loss is expected to learn the dialog model to generate utterances in the polite style when the intimacy is low and in the casual style when the intimacy is high.

Training Objective Eq. (6) shows the total loss, which is a weighted sum of the two losses concerning a style (L_w^{in} and L_s^{in}) and a general response loss (L_{NLL}).

$$L = \beta_w \cdot L_w^{in} + \beta_s \cdot L_s^{in} + \beta_{NLL} \cdot L_{NLL} \tag{6}$$

β_w, β_s, and β_{NLL} are hyperparameters representing the weight of each loss.

3.3. Additional Input

In addition to incorporating the information of user's intimacy into the loss functions, the user's level of intimacy is explicitly given in an input to the dialog model. Specifically, the level of intimacy is identified by $P(I|X_u)$, and then the intimacy label is added to the input as follows:

- When I=low : `<l> <s> context </s>`
- When I=high : `<h> <s> context </s>`

`<l>` and `<h>` are special tokens indicating the low and high intimacy classes, respectively. `<s>` and `</s>` are special tokens indicating the beginning and end of the dialog context. This additional input allows the dialog model to generate responses in an appropriate style that matches the identified level of intimacy.

3.4. Sampling and Ranking

To enhance the ability of the dialog model to generate appropriately styled utterances, the sampling-and-rank decoding strategy [12] is employed as in STYLEDGPT. First, the dialog model generates N candidate responses using top-k sampling. Next, a style score and a content score are calculated for each candidate response, Y_i, to assess the quality of Y_i. The candidate responses are then re-ranked by the weighted sum of these scores, and the response with the highest score is chosen as the final output.

The style score $\text{Score}_{style}(Y_i)$ is a weighted sum of the style probabilities of Y_i, as in Eq. (7). The weights are the probabilities of the low and high intimacy predicted by the history of the user's utterances X_u. A greater style score indicates that a response is generated in the polite (or casual) style and the user's level of intimacy is low (or high).

$$\text{Score}_{style}(Y_i) \overset{def}{=} p(I\text{=low}|X_u) \cdot p(S\text{=polite}|Y_i) + p(I\text{=high}|X_u) \cdot p(S\text{=casual}|Y_i) \tag{7}$$

The content score $\text{Score}_{content}(Y_i)$ is defined as the probability that the dialog model $P(Y|X)$ outputs the response candidate Y_i when the dialog context X is an input, as shown in Eq. (8). This score evaluates the relevance of Y_i to X.

$$\text{Score}_{content}(Y_i) \overset{def}{=} P(Y_i|X) \tag{8}$$

The final score $\text{Score}(Y_i)$ is defined as Eq. (9). The hyperparameter ω determines the relative weighting of the two scores.

$$\text{Score}(Y_i) \overset{def}{=} (1-\omega) \cdot \text{Score}_{style}(Y_i) + \omega \cdot \text{Score}_{content}(Y_i) \tag{9}$$

4. Experiments

4.1. Datasets

Dialog Corpus with Intimacy Level Our in-house dialog corpus annotated with intimacy labels was used to evaluate the proposed method. This corpus consists of recorded and transcribed dialogs of approximately ten minutes, conducted between two speakers. For each dialog, the intimacy labels of each of the two speakers to his/her dialog partner are annotated on a five-point scale. The statistics of the corpus are as follows: the number of subjects who participated in the conversations is 19, the number of conversations is 54, and the total number of utterances is 6,984. Hereafter, we refer to this corpus as the "Japanese Intimacy Dialog Corpus" or "JID corpus" for short.

The 54 dialogs in the JID corpus were 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 an instance of response. One speaker in the corpus was designated as the system and the other as the user to extract a dialog context and response. The first $n \times 2$ utterances and the next utterance in a dialog were extracted as (X, S_{n+1}). This procedure was then repeated, with the utterance shifted one by one, to obtain multiple instances of responses. Finally, 4,032, 921, and 1,284 instances of responses were obtained as the training, validation, and test data, respectively.

We also used this corpus to train an intimacy estimation model. Let $X_u = \{U_1, \cdots, U_n\}$ be the user's utterance extracted from the dialog context X in an instance of response, and let I be the intimacy label for the dialog. The intimacy label was designated as "low" when the corresponding score in the JID corpus was 1 or 2, or "high" when the value was 3, 4, or 5. The intimacy estimation model, $P(I|X_u)$, is a binary classification model that takes X_u as input and estimates the intimacy label I. The model was trained using samples (X_u, I) in the training and validation data and its performance was evaluated using the test data.

Style Corpus Two style corpora are required to train style language models and a style discrimination model: D^{po}_{style} and D^{ca}_{style}. The KeiCO corpus [9] was used as D^{po}_{style}. This corpus contains utterances using various types of honorific expressions in Japanese. Besides, D^{ca}_{style} was constructed by extracting utterances from conversations between speakers who know each other in the BTSJ corpus [24]. D^{po}_{style} contains 10,007 utterances, while D^{ca}_{style} contains 13,351 utterances.

To train the polite and the casual style language model, $P_{po}(T)$ and $P_{ca}(T)$, all utterances in D^{po}_{style} and D^{ca}_{style}, respectively, were utilized. To train the style discrimination model $P'(S|T)$, a total of 23,248 utterances were used, comprising 9,957 utterances in D^{po}_{style} and 13,301 utterances in D^{ca}_{style}. The remaining 100 utterances (50 utterances each) were used to evaluate the style discrimination model.

4.2. Experimental Setting

The following methods, including our proposed methods, were compared in the experiment.

- **DialoGPT** [22] is the dialog model based on GPT-2 [23], which has been pre-trained using a large amount of dialog data.
- **S-GPT**$_{po}$ is a STYLEDGPT that always generates polite-style responses.
- **S-GPT**$_{ca}$ is a STYLEDGPT that always generates casual-style responses.
- **Rule**$_{auto}$ is a method to control the style by heuristics. A response is generated by S-GPT$_{po}$ when the intimacy estimation model identifies the user's level of intimacy as low, and by S-GPT$_{ca}$ when it is high.
- **Rule**$_{gold}$ switches between S-GPT$_{po}$ and S-GPT$_{ca}$ based on the ground-truth label of the user's intimacy.
- **I-S-GPT**$_{auto}$ is Intimacy-aware STYLEDGPT, our proposed method.
- **I-S-GPT**$_{gold}$ is our proposed method, in which the ground-truth intimacy label is used instead of an estimate based on the intimacy estimation model.

If the performance of the intimacy estimation model is inadequate, misclassification of the level of intimacy may prevent the learning of the stylized dialog model. To verify the effectiveness of our approach to control the style of response in terms of intimacy, **I-S-GPT**$_{gold}$ was also evaluated. It can be regarded as an ideal system that always correctly estimate the user's intimacy. In this method, in Eq. (4) and (5), the probability of the level of intimacy was approximated by the five-point intimacy score (IS) in the JID corpus as $p(I\text{=low}|X_u) \simeq 1 - \frac{IS}{5}$ and $p(I\text{=high}|X_u) \simeq \frac{IS}{5}$. The additional input described in subsection 3.3 was also given by the ground-truth intimacy score, that is, <l> is added when IS is between 1–2, while <h> is added when IS is between 3–5.

A method using a Large Language Model (LLM) for style-controlled generation can be considered as a baseline. However, when a prompt is provided to ChatGPT to guess the user's level of intimacy and respond in an appropriate style, the generated responses are almost always polite. Therefore, prompting-based LLM is not included in this experiment.

4.3. Implementation Details

Style Language Model and Discrimination Model The style language models $P_{po}(T)$ and $P_{ca}(T)$ were obtained by fine-tuning GPT-2. The architecture of the style language model consists of an embedding layer, a transformer module, and a decoding layer of GPT-2. The pre-trained model was japanese-gpt2-medium[1], which had been trained on a large-scale Japanese dialog dataset. The learning rate was set to 5e^{-4}, the batch size to 4, and the epoch to 20. The Adam optimizer [25] was used to fine-tune the model.

The style discrimination model $P'(S|T)$ was also obtained by fine-tuning the GPT-2 model. The architecture of the style discrimination model consists of an embedding layer, a transformer module, and a classification layer of GPT-2. The same pre-trained model used to train the style language model was fine-tuned using the Adam optimizer with the same hyperparameters. The style discrimination model was evaluated using the 100 utterances not used for training. Its accuracy was 64%.

Intimacy Estimation Model Bidirectional Encoder Representations from Transformers (BERT) [26] was used to train the intimacy estimation model. The BERT base Japanese[2], which had been trained on Japanese Wikipedia and Japanese CC-100, was used as a pre-trained model. This BERT model was fine-tuned using the JID corpus. As for the hyperparameters, the learning rate was set to 5e^{-6}, the batch size to 1, and the epoch to 10. The Adam optimizer was used to fine-tune the model. The accuracy of the intimacy estimation model on the test set was 69%.

The low accuracy indicates that intimacy estimation is a difficult task. Our error analysis shows that there are few indicative words that are highly related to the speaker's intimacy. For example, in the sentiment analysis task, "pleasant" and "happy" are indicative words for positive emotions, and "sad" and "unhappy" are ones for negative emotions. However, such indicative words are rare in the intimacy estimation task. Another possible reason for the poor performance is the lack of training data. One of the possible directions is to apply semi-supervised learning to compensate for small amounts of labeled data with large amounts of unlabeled data.

Dialog Model The dialog model described in subsection 4.2 was obtained by fine-tuning GPT-2. The same pre-trained model[1] used for training the style language models was utilized for fine-tuning the dialog model. As for the hyperparameters, the learning rate was set to 1e^{-18}, the batch size to 1, and the epoch to 10. The Adam optimizer was used to fine-tune the model.

The parameters β_w, β_s, and β_{NLL} in Eq. (6) were set to 0.45, 0.45, and 0.1, respectively. These values were optimized on the validation data according to the StyCor criterion, which will be described in §4.4. As for the sampling-and-rank decoding strategy, the hyperparameters were set to the same values as those used in STYLEDGPT [19], specifically k to 40, N to 50, and ω in Eq. (9) to 0.5.

The length of a dialog context $X = \{S_1, U_1, \cdots, S_n, U_n\}$ was set to 8, i.e., the parameter n was set to 4. In the preliminary experiment to evaluate the intimacy estimation model, the accuracy of the model was measured for different values of n. The highest accuracy was obtained when $n = 4$.

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

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

Table 1
Results of Automatic Evaluation

Methods	Relevance					Diversity		Style
	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	ROUGE-L	Dist-1	Dist-2	StyCor
DialoGPT	0.0798	0.0110	**0.445**	**0.0617**	**0.0400**	**0.674**	**0.915**	0.115
S-GPT$_{po}$	0.0927	0.0118	0.393	0.0439	0.0244	0.648	0.897	0.0700
S-GPT$_{ca}$	**0.0933**	**0.0128**	0.392	0.0556	0.0274	0.643	0.894	0.0602
Rule$_{auto}$	0.0727	0.0082	0.428	0.0501	0.0195	0.666	0.910	0.109
Rule$_{gold}$	0.0739	0.0078	0.432	0.0477	0.0327	0.669	0.912	0.161
I-S-GPT$_{auto}$	0.0894	0.0103	0.372	0.0506	0.0230	0.660	0.900	0.103
I-S-GPT$_{gold}$	0.0715	0.0078	0.414	0.0455	0.0271	0.666	0.902	**0.366**

4.4. Evaluation Criteria

Both automatic and human evaluations were carried out to access responses generated by various methods.

Automatic Evaluation In automatic evaluation, the quality of the generated responses was evaluated from three perspectives: relevance, diversity, and style. The relevance was measured by BLEU [27] and ROUGE [28]. Specifically, the similarity between a generated response and a ground-truth response was evaluated using BLEU-1, BLEU-2, ROUGE-1, ROUGE-2, and ROUGE-L. The diversity was measured by Distinct-1 (Dist-1) and Distinct-2 (Dist-2), following the experiment by Li et al. [18]. The style was evaluated using "Style Correlation" (StyCor). The StyCor metric is defined as the correlation between the probability of the casual style $p(S\text{=casual}|Y)$ and the ground-truth level of the intimacy[3]. This correlation is high when both the predicted probability of the casual style and the intimacy level are high, or both are low (i.e., the probability of the polite style is high and the intimacy is low). It evaluates the extent to which the dialog model can control the style so that it generates a response in the casual (or polite) style when the user's level of intimacy is high (or low).

Human Evaluation The quality of the generated responses was evaluated by human subjects. A hundred instances of responses were randomly chosen from the test set in the JID corpus. For each instance, responses were generated using the methods described in subsection 4.2 against the dialog context X. The responses were then evaluated by the subjects according to the following three criteria:

- 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?

The quality of responses was evaluated by assigning a score of 3 (appropriate), 2 (neutral), or 1 (inappropriate) for each of the three perspectives. Ten native Japanese speakers participated in the human evaluation. The inter-annotator agreement was measured using Fleiss's kappa [29].

5. Results

5.1. Results of Automatic Evaluation

Table 1 shows the results of the automatic evaluation. The bold indicates the best system for each criterion. The StyCor of our proposed method using ground-truth intimacy labels, I-S-GPT$_{gold}$, was 0.366. It significantly outperformed the other baseline methods. Especially, the StyCor of I-S-GPT$_{gold}$ was much better than that of the rule-based method, Rule$_{gold}$, which naively altered the polite and

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

Table 2
Results of Human Evaluation. * means $p < 0.05$. ** means $p < 0.01$.

Model	Style Control			Relevance			Fluency		
	Score	κ	p	Score	κ	p	Score	κ	p
DialoGPT	1.98	0.26	$5e^{-5**}$	1.51	0.22	0.15	2.16	0.39	0.03*
S-GPT$_{po}$	2.08	0.18	$3e^{-4**}$	1.50	0.23	0.12	2.32	0.39	0.64
S-GPT$_{ca}$	2.05	0.19	$1e^{-3**}$	1.51	0.26	0.14	2.27	0.34	0.32
Rule$_{gold}$	2.22	0.11	0.27	1.52	0.26	0.20	2.23	0.37	0.14
I-S-GPT$_{gold}$	2.29	0.13	–	1.62	0.28	–	2.36	0.35	–

casual style generation by heuristics. These results indicated that our proposed method was superior at generating stylized responses based on the user's level of intimacy. When the user's intimacy was estimated, however, the StyCor of I-S-GPT$_{auto}$ was 0.103, which was better than STYLEDGPT but worse than DialoGPT. The poor StyCor of I-S-GPT$_{auto}$ may be due to the low accuracy (69%) of the intimacy estimation model. It was also supported by the large difference between I-S-GPT$_{auto}$ and I-S-GPT$_{gold}$. Our proposed method is highly dependent on the performance of the intimacy estimation model.

As for the relevance, S-GPT$_{ca}$ achieved the best BLEU, while DialogGPT achieved the best ROUGE. Our methods I-S-GPT$_{auto}$ and I-S-GPT$_{gold}$ were slightly worse for BLEU and obviously worse for ROUGE than the best system, but comparable to other baselines. As for the diversity, no significant difference of Dist-1 and Dist-2 was observed between the methods. From these results, it was found that the outstanding ability of the pre-trained dialog model (DialoGPT) to produce relevant and diverse responses was not remarkably damaged by incorporating the techniques of style control. Besides, no significant difference was found in relevance and diversity between I-S-GPT$_{auto}$ and I-S-GPT$_{gold}$.

5.2. Results of Human Evaluation

The automatic evaluation revealed that the StyCor scores of the methods that automatically estimated the level of intimacy (I-S-GPT$_{auto}$ and Rule$_{auto}$) were insufficiently high. These two methods were excluded from the human evaluation process to reduce the burden on the annotators.

Table 2 shows the results of the human evaluation. The "Score" column indicates the average of scores assigned by the ten annotators. The "κ" column represents Fleiss's κ, which indicates the agreement of scores between annotators. We also used Welch's test to verify whether there was a significant difference in scores between I-S-GPT$_{gold}$ and other methods. The "p" column shows the p-value associated with this statistical test.

Style Control The proposed method, I-S-GPT$_{gold}$, achieved the highest score for style control. The p-values indicated that I-S-GPT$_{gold}$ was significantly better than the other methods, except for Rule$_{gold}$. These results demonstrated that our proposed method was capable of generating responses in a more appropriate style. Rule$_{gold}$ was the second-best method, and both I-S-GPT$_{gold}$ and Rule$_{gold}$ were designed to control the style according to the level of intimacy. This confirms the validity of our approach to consider the user's level of intimacy to use polite and casual styles appropriately. However, the κ for style control was 0.13, indicating that the inter-annotator agreement was relatively low.

Relevance Although I-S-GPT$_{gold}$ was worse than the other methods in the automatic evaluation of relevance (as shown in Table 2), it achieved the highest score for relevance in the human evaluation. Nevertheless, no significant difference was observed. At least, the ability of the proposed method to generate responses relevant to the dialog context was comparable to that of the other baselines.

Fluency As with the relevance score, the average score for fluency was the highest for the proposed method. However, a significant difference was only found between DialoGPT and I-S-GPT$_{gold}$. The κ for fluency was higher than that for style control and relevance, indicating that the annotators exhibited greater consistency in evaluating the fluency of the responses.

Table 3
Average Time of Response Generation Per Utterance (seconds)

DialoGPT	**S-GPT**$_{po}$	**S-GPT**$_{ca}$	**Rule**$_{auto}$	**I-S-GPT**$_{auto}$
3.661	4.162	4.465	5.284	5.111

Computational Time Table 3 shows a comparison of the average time required for response generation per utterance across all test samples. A server with NVIDIA RTX A6000 48GB is used for the time measurements. DialoGPT exhibited the shortest generation time, followed by S-GPT, I-S-GPT, and Rule. S-GPT takes more time than DialoGPT due to the additional sampling and ranking strategy. In addition, I-S-GPT and Rule are slower than S-GPT because they require additional processing for the intimacy estimation.

6. Conclusion

This paper proposed a novel method of controlling the speech style of a dialog system according to the user's level of intimacy with the dialog system. Based on the PLM, which was a good dialog model, two loss functions were proposed to fine-tune it to generate responses in an appropriate style. In addition, the special token indicating the user's level of intimacy was added to the input of the dialog model. The results of automatic and human evaluations demonstrated that our proposed method outperformed the baseline for style control, indicating that the method could generate responses in a polite style when intimacy was low and a casual style when intimacy was high.

In the experiments, the accuracy of the intimacy estimation model was low, which caused a considerable decrease in the performance of the dialog model that used this intimacy estimation model. In the future, by improving the intimacy estimation model, we will enhance the style control ability of the dialog system under conditions where the ground-truth intimacy labels are not used.

It is our position that the study will not give rise to any significant ethical concerns. Our approach only controls speech styles according to the internal state of a user, and it does not introduce or exacerbate any ethical or social bias in a dialog system.

References

[1] C. Khatri, A. Venkatesh, B. Hedayatnia, R. Gabriel, A. Ram, R. Prasad, Alexa prize — state of the art in conversational AI, AI Magazine 39 (2018) 40–55. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2810. doi:10.1609/aimag.v39i3.2810.

[2] R. Higashinaka, K. Funakoshi, M. Inaba, Y. Tsunomori, T. Takahashi, R. Akama, Dialogue System Live Competition: Identifying Problems with Dialogue Systems Through Live Event, Springer Singapore, 2021, pp. 185–199.

[3] E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe, et al., The second conversational intelligence challenge (convai2), in: The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations, Springer, 2020, pp. 187–208.

[4] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, et al., Conversational AI: The science behind the alexa prize, arXiv preprint arXiv:1801.03604 (2018).

[5] R. Wardhaugh, J. M. Fuller, An introduction to sociolinguistics, John Wiley & Sons, 2021.

[6] E. Hovy, Generating natural language under pragmatic constraints, Journal of Pragmatics 11 (1987) 689–719. URL: https://www.sciencedirect.com/science/article/pii/0378216687901093. doi:https://doi.org/10.1016/0378-2166(87)90109-3.

[7] M. Silverstein, Indexical order and the dialectics of social life, Language & Communication 23 (2003) 193–229. doi:10.1016/S0271-5309(03)00013-2.

[8] N. Aapakallio, Understanding Through Politeness – Translations of Japanese Honorific Speech to Finnish and English, University of Eastern Finland, 2021.

[9] M. Liu, I. Kobayashi, Construction and validation of a Japanese honorific corpus based on systemic functional linguistics, in: J. Sälevä, C. Lignos (Eds.), Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 19–26. URL: https://aclanthology.org/2022.dclrl-1.3.

[10] Y. Kageyama, Y. Chiba, T. Nose, A. Ito, Improving user impression in spoken dialog system with gradual speech form control, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 235–240. URL: https://aclanthology.org/W18-5026. doi:10.18653/v1/W18-5026.

[11] T. Niu, M. Bansal, Polite dialogue generation without parallel data, Transactions of the Association for Computational Linguistics 6 (2018) 373–389. URL: https://aclanthology.org/Q18-1027.

[12] X. Gao, Y. Zhang, S. Lee, M. Galley, C. Brockett, J. Gao, B. Dolan, Structuring latent spaces for stylized response generation, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1814–1823. URL: https://aclanthology.org/D19-1190. doi:10.18653/v1/D19-1190.

[13] Q. Zhu, W. Zhang, T. Liu, W. Y. Wang, 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), Association for Computational Linguistics", Bangkok, Thailand, 2021, pp. 4391–4401.

[14] Y. Zheng, Z. Chen, R. Zhang, S. Huang, X. Mao, M. Huang, Stylized dialogue response generation using stylized unpaired texts, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, Online, 2021, pp. 14558–14567.

[15] A. Tsai, S. Oraby, V. Perera, J.-Y. Kao, Y. Du, A. Narayan-Chen, T. Chung, D. Hakkani-Tur, Style control for schema-guided natural language generation, in: A. Papangelis, P. Budzianowski, B. Liu, E. Nouri, A. Rastogi, Y.-N. Chen (Eds.), Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, Association for Computational Linguistics, Online, 2021, pp. 228–242. URL: https://aclanthology.org/2021.nlp4convai-1.21. doi:10.18653/v1/2021.nlp4convai-1.21.

[16] S. Saha, S. Das, R. Srihari, Stylistic response generation by controlling personality traits and intent, in: B. Liu, A. Papangelis, S. Ultes, A. Rastogi, Y.-N. Chen, G. Spithourakis, E. Nouri, W. Shi (Eds.), Proceedings of the 4th Workshop on NLP for Conversational AI, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 197–211. URL: https://aclanthology.org/2022.nlp4convai-1.16. doi:10.18653/v1/2022.nlp4convai-1.16.

[17] Q. Sun, C. Xu, H. Hu, Y. Wang, J. Miao, X. Geng, Y. Chen, F. Xu, D. Jiang, Stylized knowledge-grounded dialogue generation via disentangled template rewriting, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, Washington, USA, 2022, pp. 3304–3318. doi:10.18653/v1/2022.naacl-main.241.

[18] J. Li, Z. Zhang, X. Chen, D. Zhao, R. Yan, Stylized dialogue generation with feature-guided knowledge augmentation, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Sentosa Gateway, Singapore, 2023, pp. 7144–7157. doi:10.18653/v1/2023.findings-emnlp.475.

[19] Z. Yang, W. Wu, C. Xu, X. Liang, J. Bai, L. Wang, W. Wang, Z. Li, StyleDGPT: Stylized response generation with pre-trained language models, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1548–1559. URL: https://aclanthology.org/2020.findings-emnlp.140. doi:10.18653/v1/2020.findings-emnlp.140.

[20] M. Skowron, S. Rank, M. Theunis, J. Sienkiewicz, The good, the bad and the neutral: affective

profile in dialog system-user communication, in: Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part I, ACII'11, Springer-Verlag, Berlin, Heidelberg, 2011, p. 337–346.

[21] S. D'mello, A. Graesser, Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back, ACM Trans. Interact. Intell. Syst. 2 (2013). URL: https://doi.org/10.1145/2395123.2395128. doi:10.1145/2395123.2395128.

[22] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, B. Dolan, DIALOGPT : Large-scale generative pre-training for conversational response generation, in: A. Celikyilmaz, T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 270–278. URL: https://aclanthology.org/2020.acl-demos.30. doi:10.18653/v1/2020.acl-demos.30.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[24] M. Usami (Ed.), BTSJ-Japanese Natural Conversation Corpus with Transcripts and Recordings (March 2021), National Institute for Japanese Language and Linguistics, Japan, 2021.

[25] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980 (2014). URL: https://api.semanticscholar.org/CorpusID:6628106.

[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[27] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[28] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[29] J. L. Fleiss, C. Jacob, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement 33 (1973) 613–619. URL: https://cir.nii.ac.jp/crid/1360855569674739072. doi:10.1177/001316447303300309.