

Emotion-Aware Chatbot: Sentiment Analysis and Adaptive Response Generation

Kousuke Ohno
AAIL, Faculty of Data Science
Musashino University
Tokyo, Japan
S2422037@stu.musashino-u.ac.jp

Virach Sornlertlamvanich
AAIL, Faculty of Data Science
Musashino University
Tokyo, Japan
virach@musashino-u.ac.jp

Titipakorn Prakayaphun
AAIL, Faculty of Data Science
Musashino University
Tokyo, Japan
titipakorn@musashino-u.ac.jp

Thatsanee Charoenporn
AAIL, Faculty of Data Science
Musashino University
Tokyo, Japan
thatsanee@ds.musashino-u.ac.jp

Abstract—Recent advances in large-scale pre-trained language models have enabled dialogue systems to generate grammatically and semantically natural responses. However, existing chatbots often fail to produce emotionally appropriate and empathetic responses, resulting in mechanically perceived interactions. To address this issue, this paper proposes an emotion-controllable Japanese dialogue generation system that integrates emotion control directly into the generation process. The proposed system consists of a Japanese pre-trained language model (rinna Japanese GPT-2), a Plug and Play Language Model (PPLM)-based Bag-of-Words induction module, and a dialogue control interface that allows users to explicitly select emotion modes. By applying gradient-based induction to latent representations without retraining the base model, the system flexibly controls emotional nuances while preserving natural Japanese generation quality. Furthermore, emotions are treated as goal-oriented control elements that can be dynamically selected according to conversational purpose. This design enables more human-like, flexible, and empathetic dialogue compared to conventional sentiment-responsive chatbots.

Keywords—emotion control; PPLM; chatbot; BoW; sentiment selection

I. INTRODUCTION

Recent advances in neural language models have significantly improved the performance of chatbots in tasks such as information retrieval and question answering. Dialogue systems built upon large-scale pre-trained language models are now capable of generating grammatically and semantically natural responses, marking substantial progress in open-domain conversation systems [1]. Despite these advancements, many existing chatbots still struggle to reproduce the emotional richness and empathy inherent in human-to-human conversations [2].

In addition, they suffer from hallucinations, which require semantic remedies such as those proposed in Semantically Aware Reasoning (SAR) algorithm [3]. Although generated responses may appear fluent and coherent, they often lack emotional nuance and empathetic expressions that align with the speaker's feelings and conversational context, resulting in interactions that feel mechanical and impersonal.

One major reason for this limitation lies in the architectural design of current dialogue systems. Most modern chatbots rely on pre-trained language models and generate responses based on Maximum Likelihood Estimation (MLE) [4]. While this framework excels at producing statistically plausible responses from large-scale corpora, it does not provide explicit mechanisms for controlling abstract attributes such as emotion or attitude. As a result, even when users express emotions such as anxiety, joy, or frustration, these systems frequently fail to generate responses that appropriately reflect or accommodate such emotional states.

To address this issue, previous research on emotion-aware dialogue systems has often employed approaches that first estimate an emotion label from the input text and then generate or retrieve a predefined response associated with that label [5]. However, this strategy tends to produce rigid and repetitive emotional expressions, making it difficult to adapt flexibly to conversational flow or user intent. Consequently, such systems have not yet achieved the level of emotional adaptability observed in human conversations, where emotional tone is adjusted dynamically according to context and purpose [6].

These limitations indicate that merely improving response fluency is insufficient for achieving emotionally intelligent dialogue systems. Instead, emotion must be incorporated into the generation process itself as a controllable factor. Recent studies have suggested that flexible control of attributes

during generation, rather than post hoc modification or response selection, is a promising direction for overcoming the emotional rigidity of conventional systems [7].

Motivated by this perspective, this paper proposes a dialogue generation system designed to enable emotion-controllable Japanese dialogue without retraining the underlying language model. The proposed system integrates a pre-trained Japanese language model with an external emotion induction mechanism and a dialogue control interface, allowing users to explicitly specify emotional intent according to conversational goals. By separating the base language model from emotion control mechanisms, the system aims to preserve natural Japanese generation while flexibly controlling emotional nuance, thereby contributing to more human-like and empathetic conversational interactions.

II. RELATED WORK

A. Large-Scale Pre-trained Language Models for Dialogue

Large-scale pre-trained language models have significantly advanced dialogue generation by enabling grammatically and semantically coherent responses. Models such as GPT-2 and GPT-3 demonstrate strong capabilities in open-domain dialogue generation by leveraging massive text corpora and autoregressive generation frameworks [1]. These models, however, primarily optimize response likelihood through Maximum Likelihood Estimation (MLE), which focuses on statistical plausibility rather than controllability of abstract attributes such as emotion or attitude [4]. As a result, while surface-level fluency has improved, emotional appropriateness and empathetic response generation remain limited.

B. Emotion-Aware and Sentiment-Based Dialogue Systems

To address emotional inadequacies, earlier emotion-aware dialogue systems often incorporated sentiment analysis modules that classify user input into discrete emotion labels, followed by generating or selecting predefined responses associated with those labels [5]. While such approaches enable basic emotional responsiveness, they suffer from several limitations, including rigid emotional expression, poor adaptability to dialogue context, and susceptibility to sentiment misclassification. Moreover, these systems often fail to reflect the user's true communicative intent, especially in cases involving ambiguity or context-dependent emotion [2, 6].

C. Controllable Text Generation and Attribute Control

Recent research has explored controllable text generation techniques that enable manipulation of specific attributes without retraining language models. Plug and Play Language Models (PPLM) introduce a framework that applies gradient-based updates to latent representations during generation, allowing external control over attributes such as sentiment or

topic [7]. Unlike fine-tuning approaches, PPLM preserves the original language model parameters while enabling flexible attribute induction. This method has been shown to be effective for sentiment control using Bag-of-Words (BoW) guidance, making it suitable for interactive dialogue scenarios.

D. Goal-Oriented Emotion Control in Dialogue

Beyond static emotion control, recent dialogue research emphasizes the importance of goal-oriented and context-aware emotion modulation. Human conversations naturally adjust emotional tone according to conversational goals, such as providing emotional support or conveying factual information [8]. Incorporating such adaptive emotion selection into dialogue systems has been shown to improve user satisfaction and perceived naturalness [9]. Vocabulary-level control methods further enhance expressive diversity by allowing fine-grained emotional adjustments aligned with user intent [10].

E. Main Contribution of This Work

In contrast to conventional sentiment-responsive chatbots, this work adopts a user-driven emotion specification strategy combined with PPLM-based BoW induction. By separating the base language model, emotion induction mechanism, and dialogue control interface, the proposed system enables flexible, goal-oriented emotion control without retraining [1]. This design bridges the gap between fluent language generation and interactive emotional adaptability, contributing to more human-like Japanese dialogue generation [2].

III. LIMITATION OF LLM BASED CHATBOT

In recent years, advancements in neural language models have enabled chatbots to demonstrate high performance in tasks such as information retrieval and question answering. In particular, dialogue systems based on large-scale pre-trained language models have made remarkable progress in generating grammatically and semantically natural responses. However, many existing chatbots face the challenge of insufficiently reproducing the emotional exchanges characteristic of human-to-human conversations.

As shown in Fig. 1, it has been pointed out that generated responses, although formally natural, often lack empathic expressions and emotional nuances appropriate to the speaker's feelings and the conversational context, resulting in a mechanical impression. The generated responses may appear formally natural but often lack empathetic expressions and emotional nuances, resulting in a mechanical impression [2].

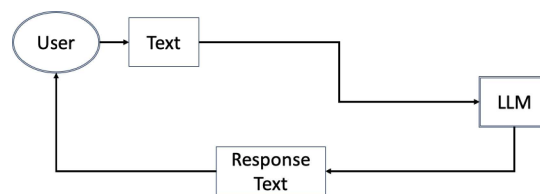


Figure 1. Current utilization of LLM in conversational chatbot

Underlying this problem are the limitations of the response generation structures adopted by mainstream chatbots. Many dialogue systems are based on pre-trained language models and generate responses to input sentences using a framework based on Maximum Likelihood Estimation (MLE) [1]. While this structure excels at generating statistically plausible responses from large-scale corpora, it lacks mechanisms to explicitly control abstract attributes like emotion or attitude during the generation process (This reflects the problem structure identified in prior research: “MLE-based generation frameworks lack explicit control mechanisms for emotion or attitude”). Consequently, it has been reported that even when users express emotions like anxiety or joy, generating appropriately empathetic emotional responses is difficult (“users express emotions like anxiety or joy, but the system struggles to generate appropriately empathetic responses” reflects the problem structure in conventional research).

Furthermore, conventional emotion-response chatbots have widely employed a technique where emotion labels are estimated from input sentences, and a predefined response corresponding to that label is returned [5]. However, this approach suffers from the problem that emotional expressions tend to become fixed, making it difficult to flexibly adjust emotions according to the user's dialogue intent or the flow of conversation (“template-based emotion responses are fixed and inflexible to dialogue flow or user intent” as the problem structure in prior research) [6]. Consequently, it has been pointed out that they fail to achieve the naturalness of human dialogue, where emotional tone changes according to purpose and context (“emotional tone changes according to purpose and context, which conventional bots fail to reproduce” – the limitation structure in previous research) [6, 7].

As described above, current chatbots lack structural designs that incorporate emotion into the generation process itself, leaving structural challenges in realizing emotional responses and empathetic dialogue (the problem structure identified in previous research: “current chatbots lack structural designs that incorporate emotion into the generation process itself”) [7]. To solve this problem, it is essential to integrate a mechanism into the dialogue system structure that allows flexible control of emotional nuance during generation, beyond merely improving response accuracy (the problem structure of previous research: “flexible emotional nuance control during generation is essential beyond response accuracy”) [7].

IV. SYSTEM OVERVIEW

This system is a dialogue generation system designed to generate emotion-controllable Japanese dialogue. It consists of three main components: a Japanese pre-trained language model (rinna Japanese GPT-2), a Bag-of-Words (BoW) induction module based on PPLM, and a dialogue control interface that selects emotion modes and adjusts generated sentence length. As shown in Fig. 2, these components are

organized to process user input, analyze sentiment, and generate emotionally controlled responses.

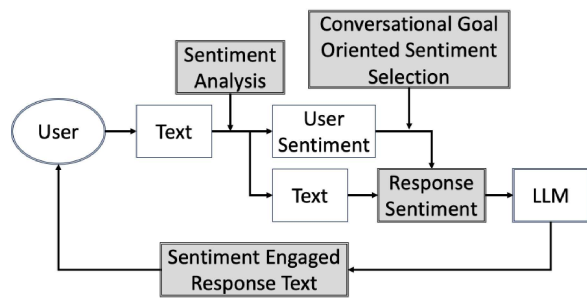


Figure 2. Proposed utilization of LLM for sentiment engaged conversation

By combining these components, the system aims to maintain the ability to generate natural Japanese responses while controlling the emotional nuances contained within the generated text.

When a user inputs a sentence and specifies an emotion mode (Positive, Neutral, or Custom), the dialogue control interface parses this control information, selects the vocabulary set corresponding to the specified emotion mode, and passes it to the BoW induction module. This ensures that the emotional information is reflected in the generation process as an explicit vocabulary set.

The BoW induction module performs gradient ascent induction on the language model’s latent state according to the PPLM framework, making the specified vocabulary more likely to be generated. PPLM is a technique that enables attribute control of generated text without updating model parameters by sequentially adjusting the latent representation during generation towards the attribute direction [7]. This induction process enables the assignment of emotional tendencies across the entire generated response.

The latent representations obtained after induction are input to the underlying Japanese GPT-2 model. However, the parameters of this model remain fixed, preserving the Japanese generation capabilities acquired through pre-training. Consequently, responses are generated and presented to the user while maintaining grammatically and semantically natural expressions, with only the emotional nuances flexibly controlled.

As described above, by adopting a structure that clearly separates the base language model, attribute induction mechanism, and dialogue control mechanism, this system achieves interactive sentiment control without requiring retraining.

V. METHODS

A. Methods for Sentiment Analysis of Input Text

The sentiment analysis of input sentences in this system adopts a design approach distinct from the automatic sentiment classification methods commonly used in conventional sentiment-responsive chatbots. Many existing studies have employed a method where

emotion labels are estimated from user input sentences, and responses are generated based on these results. However, such approaches are susceptible to misclassification and have been noted to not always accurately reflect the user's conversational intent [2].

In contrast, this system does not perform automatic emotion estimation on the input text itself. Instead, it adopts a method where the user explicitly specifies an emotion mode (Positive, Neutral, or Custom) according to the purpose and context of the interaction. That is, emotion information is not indirectly inferred from the content of the input text but is treated as control information directly provided by the user's choice.

This design avoids misrecognition caused by ambiguous expressions, sarcasm, or context-dependent emotional expressions within the input sentence. It also enables emotion control that more accurately reflects the user's own intent. Furthermore, since it does not require training or inference of emotion classification models, it contributes to simplifying system architecture and reducing computational costs. This user-driven emotion specification can be considered a practical design that complements the limitations of conventional methods that rigidly estimate emotions.

B. Methods for Introducing Emotion in Output Sentence

Emotion introduction into output sentences is achieved using a Bag-of-Words (BoW) induction method based on PPLM (Plug and Play Language Model). PPLM is proposed as a technique to control the attributes of generated sentences without retraining the model, by applying external control signals to the latent states of the language model during the generation process [7].

In this system, predefined vocabulary sets are prepared as BoW depending on the emotion mode. In Positive mode, a set of Japanese vocabulary expressing joy and positivity is selected. In Custom mode, a vocabulary set specified by the user is used. These vocabulary sets are input into the BoW induction module of PPLM, where control is applied to increase the probability of their inclusion in the generated text.

Specifically, gradient ascent updates are iteratively applied to the latent state during generation, increasing the probability of specified vocabulary appearing. During this process, the parameters of the underlying Japanese GPT-2 model remain fixed, preserving the grammatical and semantic Japanese generation capabilities acquired through pre-training. This design avoids the problem inherent in standard MLE-based generation methods—namely, the inability to add sentiment after generation—by embedding emotional nuance directly into the generation process itself [4].

In contrast, Neutral mode does not perform BoW induction. Responses are generated based solely on the language model's inherent generative distribution, enabling the production of objective responses free from emotional bias.

C. Goal-Oriented Emotion Selection Theory

This system treats emotion not as a fixed attribute, but as a controllable element that can be selected according to the purpose of the conversation. This design reflects the characteristic of human conversation where the intensity and tone of emotional expression are adjusted based on the topic, situation, and conversational goal, a feature whose importance has been highlighted in recent dialogue system research [8, 9].

For instance, when a user consults about anxiety or worries, selecting Positive mode generates responses more likely to include encouragement and positive vocabulary. This enables dialogue that goes beyond mere information provision, offering empathetic and emotionally supportive interaction attuned to the user's emotional state. Such emotional consideration is regarded as a crucial element in achieving natural human-computer dialogue [9].

Conversely, in dialogues primarily focused on fact-checking or explanation, using Neutral mode generates objective, information-centered responses with restrained emotional expression. Adjusting the degree of emotional expression according to the dialogue's purpose is known to contribute to efficient information transfer in collaborative dialogues [8].

Furthermore, when intending to deliberately emphasize specific topics or emotional tones, the Custom mode allows users to specify their own vocabulary sets, enabling flexible emotion control tailored to the purpose. Such vocabulary-level emotion control facilitates diverse expressions reflecting the dialogue context and user intent, serving as an effective means to expand the range of emotional expression [10].

As described above, by adopting a design that dynamically selects and controls emotions according to the conversational goal, this system aims to mitigate the challenge of rigid emotional expression common in existing chatbots and realize a more human-like, flexible conversational experience.

VI. EVALUATION

This evaluation verifies whether the proposed system simultaneously satisfies the following criteria: (a) the ability to reflect emotions in responses according to the specified emotion mode, (b) preservation of the language quality of the base model, and (c) stability and diversity of generation. The evaluation design is constructed based on the theoretical premises of PPLM's effectiveness in achieving attribute control without retraining [7] and the conversational applicability of style and attribute injection via BoW induction [4, 11].

A. Experimental Setup

Evaluation was conducted by inputting three sentences per user input across three emotion modes (Positive / Negative / Neutral) and observing the system's generated responses. The base model used was rinna Japanese GPT-2, with only gradient

induction applied to the latent state without updating model parameters.

This configuration strictly adheres to the “plug-and-play” design [7], which performs attribute control without disrupting the pre-trained language model distribution.

B. Results and Analysis

(1) Positive Mode

In Positive mode, we confirmed that injecting a BoW vocabulary set (e.g., “excited,” “happy,” “wonderful,” “looking forward”) into the latent state via PPLM induction resulted in responses biased toward a positive overall tone.

Input: 「今日は天気が良いですね」

“The weather is nice today, isn't it?”

Response: 「先日、ご近所の方の畑を案内してきました(^)…」

“The other day, I showed a neighbor around their field (^)…”

While the response deviates from the weather topic and reduces conversational relevance, the emoji and softer tone contribute to a positive conversational style. This aligns with PPLM's behavior [7], where lexical induction influences the overall emotional tendency of the sentence.

Input: 「新しい仕事が始まります」

“A new job is starting.”

Response: 「ワクワクするような仕事を ワクワクするような仕事を…」

“An exciting job... An exciting job…”

Repetition occurs, compromising generative diversity and fluency. This corresponds to the issue noted in [4] that excessive attribute-oriented guidance may lead to monotonous generation.

Input: 「最近調子が悪いです」

“I've been feeling off lately.”

Response: 「自分の好きなことが仕事なんだと。仕事は楽しいです…」

“My job is doing what I love. Work is fun…”

While positive vocabulary is introduced in response to the user's negative context, it does not constitute an empathetic response. This demonstrates the limitation that empathy generation cannot be guaranteed by BoW prompting alone.

(2) Negative Mode

In Negative mode, inducing BoW terms like “anxiety,” “tears,” “painful,” and “suffering” resulted in responses reflecting the overall emotional tone of the sentence. However, this approach showed a tendency toward further deterioration in information accuracy and dialogue relevance.

Input: 「今日はいい天気ですね」

“It's nice weather today, isn't it?”

Response: 「でもまだ少し蒸し暑いな…暑い日になって

きています」

“But it's still a bit muggy... It's getting hotter.”

The negative tone adjustment succeeded, making it appropriate for speaker tone control. However, its reliance on lexical prompting rather than contextual understanding aligns with PPLM behavior.

Input: 「新しい仕事が始まります」

“A new job is starting.”

Response: 「新しい仕事を始める 新しい仕事の 新しい仕事 新しい…」

“Starting a new job. A new job. A new job. A new job…”

Repetition and fragmentation are prominent, indicating compromised stability in latent state induction. This aligns with the risk of generation instability due to non-learning-based induction [11].

Input: 「明日のことを考えると」

“When I think about tomorrow…”

Response: 「不安にかられて涙が止まらなかった…でも僕のせいじゃない」

“I was overcome with anxiety and couldn't stop crying... But it's not my fault.”

The emotional tone is appropriately guided towards the Negative direction, but the output resembles soliloquy generation rather than a dialogue response.

(3) Neutral Mode

In Neutral mode, omitting BoW induction yields responses based on the model's inherent generative distribution, achieving sentence generation with low emotional bias. However, while formally natural, deviations and repetitions persist in the dialogue context.

Input: 「AIについてどう思いますか」

“What do you think about AI?”

Response: 「AIについてどう思いますか あなたが、私にあなたの感情が…」

“What do you think about AI? You, tell me your feelings…”

The sentence structure resembles a question, but the response rephrases the query back to the user.

Input: 「天気予報を教えてください」

“Please tell me the weather forecast.”

Response: 「昨日、一昨日と、雨の中、沢山の方にご来場いただき…」

“Despite the rain yesterday and the day before, many people came to visit…”

It does not fulfill the goal of providing a weather forecast, but shows no emotional bias.

Input: 「今日は何曜日ですか」

“What day is it today?”

Response: 「今日は何曜日ですか 今日は何曜日ですか こんにちは…」

“What day is it today? What day is it today?
Hello...”

Repetition and mixed greetings occur, but no bias from emotional induction is observed.

C. Discussion

The evaluation results support the objective that this system can control the overall emotional tone of responses by inducing PPLM+BoW into latent states (Hypothesis 3(2)). However, it also confirms the limitation that BoW induction alone does not guarantee consistency with dialogue relevance, empathy, or informational purpose.

This observation aligns with existing findings [4, 11] that BoW strongly contributes to style control but does not guarantee semantic coherence.

Furthermore, we observed that excessive guidance toward attribute directions can lead to repetition, monotony, and context deviation. This suggests that future improvements should include dynamic adjustment of guidance strength and the addition of reward and objective functions.

VII. CONCLUSION

This paper proposes a dialogue system that can control the emotional tone of Japanese dialogue generation without retraining, by utilizing PPLM and Bag-of-Words (BoW) induction. The proposed system is based on rinna Japanese GPT-2 and achieves attribute control without updating the base model's parameters through a plug-and-play design that gradient-induces the emotional vocabulary set into the latent state [7]. This structural separation is valid from the perspective of preserving the linguistic quality of large-scale pre-trained models [1] and aligns with the design philosophy of the Conditional Transformer, which grants controllability while maintaining the ability to generate statistically plausible responses [4].

Evaluation confirmed that in Positive/Negative modes, lexical-level sentiment orientation is reflected across the entire response [7]. However, it also revealed limitations: BoW alone struggles to guarantee semantic relevance in dialogue or empathetic responses [10]. Furthermore, it was observed that excessively strong latent state induction tends to cause generation instability, such as repetition, monotonicity, and context deviation [11]. This aligns with recent controlled generation research pointing out that non-learning attribute induction can locally distort the model's distribution, potentially compromising fluency and diversity [11].

In summary, this system supports the validity of the design hypothesis that “emotion is introduced and controlled within the generation process itself.

However, it also suggests that ensuring conversational coherence and empathy quality requires not only lexical induction but also adaptive control of the objective function, conversational understanding mechanism, and induction strength. Future challenges include: (1) integrating context understanding with emotion control, (2) implementing dynamic scheduling of prompting intensity, and (3) introducing additional control objectives considering empathy and goal consistency.

Achieving these improvements is expected to enable the realization of a dialogue system that simultaneously achieves goal-oriented and flexible emotion control resembling human interaction [2] and objective response generation free from subjectivity and emotional bias [10].

REFERENCES

- [1] T. Brown *et al.*, “Language Models are Few-Shot Learners,” *arXiv preprint arXiv:2005.14165*, 2020, doi: 10.48550/arXiv.2005.14165.
- [2] Z. Hao, H. Minlie, Z. Tianyang, Z. Xiaoyan, and L. Bing, “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory,” *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 730–738, 2018.
- [3] S. Virach, “Enhancing large language models: Alleviating knowledge deficiency with external knowledge and semantically aware reasoning (SAR),” *Knowledge and Information Systems*, Springer Nature, pp. 1–19, 2025, doi: 10.1007/s10115-025-02485-4.
- [4] N. K. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, “CTRL: A Conditional Transformer Language Model for Controllable Generation,” *arXiv preprint arXiv:1909.05858*, 2019.
- [5] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, “A Persona-Based Neural Conversation Model,” *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 994–1003, 2016.
- [6] W. Li, J. A. M. Van der Aalst, and T. Okimoto, “A Systematic Review on Artificial Intelligence-Based Multimodal Dialogue Systems Capable of Emotion Recognition,” *IEEE Access*, vol. 13, pp. 1–25, 2025.
- [7] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, “Plug and Play Language Models: A Simple Approach to Controlled Text Generation,” *arXiv preprint arXiv:1912.02164*, 2020.
- [8] R. Carter, *Using Language*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
- [9] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.
- [10] A. Balahur, R. Mihalcea, and A. Montoyo, *Computational Approaches to Subjectivity and Sentiment Analysis*. Berlin, Germany: Springer, 2014.
- [11] K. Yang and D. Klein, “FUDGE: Controlled Text Generation With Future Discriminators,” *Proc. 2021 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3511–3535, 2021.