

Time-Considerable Dialogue Models via Reranking by Time Dependency

Yuiko Tsunomori¹, Masakazu Ishihata¹, Hiroaki Sugiyama¹

¹NTT Communication Science Laboratories

{yuiko.tsunomori, masakazu.ishihata, hiroaki.sugiyama}@ntt.com

Abstract

In the last few years, generative dialogue models have shown excellent performance and have been used for various applications. As chatbots become more prevalent in our daily lives, more and more people expect them to behave more like humans, but existing dialogue models do not consider the time information that people are constantly aware of. In this paper, we aim to construct a *time-considerable* dialogue model that actively utilizes time information. First, we categorize responses by their naturalness at different times and introduce a new metric to classify responses into our categories. Then, we propose a new reranking method to make the existing dialogue model *time-considerable* using the proposed metric and subjectively evaluate the performances of the obtained time-considerable dialogue models by humans.¹

1 Introduction

In the last few years, generative dialogue models have achieved outstanding performance (Ziegler et al., 2019; Adiwardana et al., 2020; Ouyang et al., 2022; Thoppilan et al., 2022) and have been used in various applications, including search engines, recommendations, healthcare, finance, and more (Ling et al., 2023). As chatbots permeate our daily lives, more and more people expect chatbots to behave in a human-like manner. Examples of research to make chatbots more human-like include the introduction of common sense (Wang et al., 2020), empathy (Ma et al., 2020), personas (Zhang et al., 2018), and so forth. The common point among these studies is that they have achieved richer dialogue by actively utilizing not only internal information obtained through the conversation but also *external information* that does not appear in the current dialogue. On the other hand, *time information*,

¹The detailed information about our dataset is available at <https://github.com/nttcsllab/time-considerable-dialogue-model>

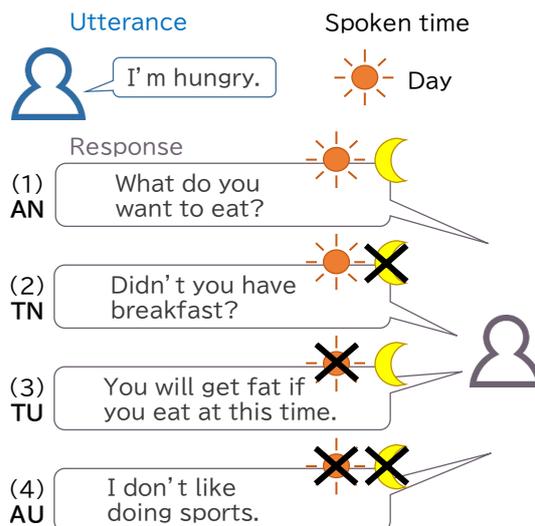


Figure 1: Conversation examples depending on time.

the most basic and important external information, still does not seem to be considered important in dialogue models.

Humans are basically always aware of time in conversation, whether explicitly or implicitly, because the *naturalness* of utterances and responses may change based on their spoken times (e.g., time of the day, day of the week, and season). Given an utterance and its spoken time, a response can be categorized into the following four types by focusing on the time variations of its naturalness:

- AN: always natural
- TN: temporarily natural at the spoken time
- TU: temporarily unnatural at the spoken time
- AU: always unnatural

Figure 1 shows examples of the above four types of responses. Assume two time periods, day and night, where responses (1)-(4) have different levels of naturalness at different times. Responses (1) and (2) ($N \triangleq AN \cup TN$) are natural to a given utterance at a spoken time, although not (3) and (4)

($\mathbf{U} \triangleq \mathbf{TU} \cup \mathbf{AU}$). On the other hand, the naturalness of (2) and (3) ($\mathbf{T} \triangleq \mathbf{TN} \cup \mathbf{TU}$) changes with the spoken time (day or night), and that of (1) and (4) ($\mathbf{A} \triangleq \mathbf{AN} \cup \mathbf{AU}$) remains unchanged. Hereafter in this paper, we refer to such categorization of responses as the *NUTA categories*.

For a dialogue model to achieve natural conversation, it is expected to generate natural responses \mathbf{N} and to avoid generating unnatural responses \mathbf{U} . If a dialogue model correctly evaluates the naturalness of responses considering the spoken time, we call it *time-aware*; otherwise, we call it *time-unaware*. Many existing dialogue models are time-unaware because they are trained on datasets *without* time information. As a result, a time-unaware dialogue model may consider \mathbf{TU} responses as natural as \mathbf{N} responses because they are natural at some time. In other words, it may generate a \mathbf{TU} response, which is inappropriate at the spoken time. The simplest way to construct a time-aware dialogue model is to train a general dialogue model with time information (Sato et al., 2017). Recently, the output of a large language model (LLM) was adjusted by giving an appropriate prompt (Lester et al., 2021; Bae et al., 2022; Liu et al., 2022), and prompting the time information to an LLM is another promising way to achieve a time-aware dialogue model. The difference between time-unaware and time-aware dialogue models is that the former considers a response as natural if it is natural at some time, while the latter does if it is natural at its spoken time; namely, time-aware dialogue models avoid generating \mathbf{TU} responses.

Prior research has shown that users’ impressions of dialogue models are improved by *actively* utilizing external information (Vinyals and Le, 2015; Li et al., 2016; Zhou et al., 2021). Referring to this fact, we assume that users’ impressions will be similarly improved by actively utilizing time information; namely, *users prefer \mathbf{TN} to \mathbf{AN}* (we empirically verify this assumption in Section 3). Under this assumption, we aim to realize a *time-considerable* dialogue model that actively outputs more \mathbf{TN} than \mathbf{AN} . The difference between time-aware and time-considerable models is that the former only considers time information to evaluate the naturalness of responses at a given spoken time (*time-aware naturalness*), while the latter actively generates responses whose naturalness varies with time. To realize time-considerable models, we need a new criterion to distinguish between \mathbf{TN}

and \mathbf{AN} and a new mechanism to generate \mathbf{TN} responses.

In this paper, we propose a new reranking method that is a post-processing method to make existing dialogue models more time-considerable. As a preliminary analysis, we first verify our assumption *users tend to prefer \mathbf{TN} to \mathbf{AN}* by human evaluation in Section 3. We next formally define the NUTA categories and propose an automatic metric for the NUTA categories in Section 4. In Section 5, we propose a new reranking method using the proposed automatic metric to make existing dialogue models time-considerable and subjectively evaluate obtained time-considerable models to verify whether our reranking method improves response qualities.

2 Related Work

Previous studies have pointed out that general dialogue models trained on large-scale datasets tend to generate neutral (bland, generic, and hackneyed) responses (Li et al., 2016; Serban et al., 2016). To tackle this issue, some dialogue models utilizing *external information* to generate interesting and informative responses have been proposed, where examples of external information include system/user persona (Zhang et al., 2018; Roller et al., 2021; Lu et al., 2022), knowledge graphs (Zhang et al., 2020), knowledge sources (Parthasarathi and Pineau, 2018; Majumder et al., 2022), interpersonal relationships (Utami and Bickmore, 2019), and situated environments (Misu, 2018). Some studies have empirically shown that generating responses specific to external information improved users’ impression of the dialogue models (Vinyals and Le, 2015; Li et al., 2016; Zhou et al., 2021). In this paper, we introduce the time information as external information and consider that \mathbf{T} responses, whose naturalness varies with time, are specific to time information. Furthermore, to verify whether the time information improves model performance as well as other external information, we evaluate the qualities of \mathbf{T} responses by human evaluations.

In this paper, we define a dialogue model as *time-aware* if the model evaluates the naturalness of responses considering their spoken times. Sato et al. (2017) proposed a time-aware dialogue model that is an encoder-decoder model based on Long-Short Term Memory (LSTM) (Zaremba et al., 2015) inspired by Johnson et al. (2017) and trained on utterance-response pairs with their timestamps ex-

tracted from Twitter. Another recent technique to achieve a time-aware dialogue model is *prompting* LLMs, where prompting is a technique to guide LLMs in generating high-quality and relevant responses by providing detailed descriptions and/or input-output examples of the target task as input (Brown et al., 2020). Various prompting-based methods for utilizing external information have been proposed, and their examples include system/user persona information (Kasahara et al., 2022; Lee et al., 2022), knowledge sources (Liu et al., 2022), and fictional character’s style (Han et al., 2022). In this paper, we aim to construct a *time-considerable* dialogue model, which actively utilizes the time information, and propose a new reranking method to make the existing dialogue models more time-considerable using time-aware naturalness represented by a time-aware dialogue model. While time-aware dialogue models only consider time to evaluate the naturalness of responses, time-considerable dialogue models actively generate responses whose naturalness varies with time.

3 Preliminary Analysis

Through this paper, we assume *users prefer TN to AN*, and in this section, we verify this assumption by human evaluations. We first constructed a NUTA dataset consisting of the tuples of utterances, their spoken times, their responses, and their NUTA categories: **AN**, **TN**, **TU**, and **AU**. Then, we conducted a subjective evaluation to determine which category of responses users found most interesting and informative.

3.1 The NUTA Dataset

We constructed a NUTA dataset, which is a collection of tuples $\mathbf{t} \triangleq \langle u, t, r, c \rangle$, where u is an utterance, $t \in \mathcal{T} \triangleq \{0, 1, \dots, 23\}$ is its spoken time in 24-hour time format, r is a response to u , and $c \in \{\mathbf{AN}, \mathbf{TN}, \mathbf{TU}, \mathbf{AU}\}$ is its NUTA category.

We first prepared a set of utterances \mathcal{U} . We extracted Japanese tweets posted between May and December 2022 with filtering rules described in Appendix A.1 and randomly selected 1,000 tweets. We manually deleted tweets containing discriminatory, violent, or other inappropriate expressions. As a result, we obtained 640 appropriate tweets and used them as utterances \mathcal{U} .

We next obtained responses to the prepared utterances \mathcal{U} by crowdsourcing, where we used

Lancers², a Japanese crowdsourcing service. We assigned one crowd worker to each utterance $u \in \mathcal{U}$ and asked them to perform the following tasks to create responses:

1. Create response $r_u^{\mathbf{AN}}$ to u that is natural at any time.
2. Select two time periods $t^{\mathbf{N}}$ and $t^{\mathbf{U}}$ ($t^{\mathbf{N}}, t^{\mathbf{U}} \in \mathcal{T}$) and create response $r_u^{\mathbf{TN}}$ to u that is natural at $t^{\mathbf{N}}$ but not at $t^{\mathbf{U}}$.

For instance, given utterance $u =$ “It seems the train is stopped,” the crowd worker selected two time periods $t = 22$ and $t' = 6$ and created two responses $r_u^{\mathbf{AN}} =$ “Really? I wonder when it will start moving.” and $r_u^{\mathbf{TN}} =$ “Wow, it’s almost the last train. I wonder what’s going to happen?”

Finally, we constructed a NUTA dataset by creating the following four tuples for each utterance $u \in \mathcal{U}$, where u' was randomly chosen from \mathcal{U} so that $u \neq u'$:

$$\begin{aligned} &\langle u, t^{\mathbf{N}}, r_u^{\mathbf{AN}}, \mathbf{AN} \rangle, & \langle u, t^{\mathbf{N}}, r_u^{\mathbf{TN}}, \mathbf{TN} \rangle, \\ &\langle u, t^{\mathbf{U}}, r_u^{\mathbf{TN}}, \mathbf{TU} \rangle, & \langle u, t^{\mathbf{N}}, r_{u'}^{\mathbf{AN}}, \mathbf{AU} \rangle. \end{aligned}$$

Thus, the dataset consists of 2,560 ($= |\mathcal{U}| \times 4$) tuples. Table 1 shows examples of four created tuples for the same utterance.

3.2 Subjective Evaluation

We conducted a subjective evaluation to determine which response category is the most interesting and informative for humans.

We introduced a new metric to measure the quality of responses considering time information. The metric is in the range $[0, 1]$ and based on the Sensibleness, Specificity, Interestingness (SSI) metric (Thoppilan et al., 2022) for evaluating responses based on context. Our metric, denoted by SSI-t, averages the following four scores:

- *Sensibleness for the utterance* (SU): If its spoken time is ignored, is the response reasonable to its utterance?
- *Sensibleness for the spoken time* (ST): If its utterance is ignored, is the response reasonable to its spoken time?
- *Specificity to time* (S): Is the response specific to any time regardless of its spoken time?
- *Interestingness* (I): Is the response interesting or informative?

²<https://www.lancers.jp/>

Utterance $u =$ It seems the train is stopped. (電車止まってるらしい..)		
Time t	Response r	Category c
22	Really? I wonder when it will start moving. (そうなんだ。いつになったら動くのかなー。)	AN
22	Wow, it's almost the last train. I wonder what's going to happen? (ええー。あと少しで終電だけど、どうなるのかな?)	TN
6	Wow, it's almost the last train. I wonder what's going to happen? (ええー。あと少しで終電だけど、どうなるのかな?)	TU
22	Understood! I'll decide on the character as soon as possible. (了解! なるべく早くキャラを決めるわ)	AU

Table 1: Four tuples $\mathbf{t} = \langle u, t, r, c \rangle$ for the same utterance. A crowd worker selected $t^{\mathbf{N}} = 22$ and $t^{\mathbf{U}} = 6$ and generated two sentences $r_u^{\mathbf{AN}}$ and $r_u^{\mathbf{TN}}$. The utterance and all responses were originally written in Japanese and translated into English by the authors.

For example, a score SU of 1.0 indicates r is a perfect response to u if t is ignored, and a score S of 0.0 indicates the naturalness of r never changes over time.

We randomly selected 25 tuples for each NUTA category (100 tuples in total). For any tuple $\mathbf{t} = \langle u, t, r, c \rangle$ and any score $SC \in \{\text{SU}, \text{ST}, \text{S}, \text{I}\}$, we asked two expert annotators, who are in-house workers specialized in annotating dialogues and have worked in their positions for at least five years, to rate SC of \mathbf{t} with either 0 or 1. We defined the SC value of \mathbf{t} as the average of two obtained rates and the SSI- t value of \mathbf{t} as the average of all SC values of \mathbf{t} . Finally, for any NUTA category c and any score $SC \in \{\text{SU}, \text{ST}, \text{S}, \text{I}, \text{SSI-}t\}$, we obtained the SC value of c by averaging those of all tuples \mathbf{t} whose categories are c . Table 2 shows four examples of tuples and their obtained values.

Table 3 shows the SSI- t scores for each NUTA category and indicates that TN achieved the highest quality (SSI- t): more specifically, the highest ST and I scores. We believe that the results support our assumption that *users prefer TN to AN*, and based on this assumption, we will propose a method to realize a time-considerable dialogue model that actively outputs more TN than AN.

4 Automatic Metric for NUTA Categories

We propose a new automatic metric for classifying the NUTA categories of given responses and experimentally show that our metric can correctly categorize the responses of the NUTA dataset.

4.1 Definition

We mathematically introduce the *time-aware naturalness* and the *time dependency* of responses and define the NUTA categories using those quantities. Let u , r , and t be an utterance, a response to u , and the time at which the conversation took place.

Time-aware naturalness We assume that the *time-aware naturalness* (TAN) of u and r at t is *implicitly* defined by conditional probability distribution $p(u, r | t) = p(u | t)p(r | u, t)$, where $p(u | t)$ and $p(r | u, t)$ indicate the TANs of u at t and r given u at t . We consider \mathbf{N} responses of the NUTA categories as natural at spoken time t ; i.e., response r is classified as \mathbf{N} (resp. \mathbf{U}) iff $p(r | u, t)$ is high (resp. low). Since it is very difficult to know the true TAN p , throughout this paper, we assume that TAN p is given as a time-aware dialogue model that allows us to evaluate $p(r | u, t)$ for any u , r , and t .

Change of naturalness Using TAN p , we define the change of (log) naturalness (CN) from t' to t as

$$\text{CN}_{t':t}(u, r) \triangleq \ln \frac{p(u, r | t)}{p(u, r | t')}, \quad (1)$$

$$\text{CN}_{t':t}(u) \triangleq \ln \frac{p(u | t)}{p(u | t')}, \quad (2)$$

$$\text{CN}_{t':t}(r | u) \triangleq \ln \frac{p(r | u, t)}{p(r | u, t')}. \quad (3)$$

By definition, the following equation must hold:

$$\text{CN}_{t':t}(u, r) = \text{CN}_{t':t}(u) + \text{CN}_{t':t}(r | u), \quad (4)$$

where the CN of conversation (u, r) can be factorized into the CNs of u and r given u . For instance, $\text{CN}_{t':t}(u, r) > 0$ holds iff the conversation (u, r) is more natural at t than at t' .

Time dependency Using the above CN, we define *time dependency* (TD) of r given u as

$$\text{TD}(r | u, t) \triangleq \max_{t^{\mathbf{U}} \in \mathcal{T}} \text{CN}_{t^{\mathbf{U}}:t}(r | u), \quad (5)$$

$$\text{TD}(r | u) \triangleq \max_{t^{\mathbf{N}} \in \mathcal{T}} \text{TD}(r | u, t^{\mathbf{N}}). \quad (6)$$

$\text{TD}(r | u, t)$, denoted by $\text{TD}@t$, is the CN of r given u from the most unnatural time $t^{\mathbf{U}}$ to spoken

Utterance u	Time t	Response r	Cat. c	SU	ST	S	I	SSI-t
If gummies are within reach, I can't stop eating them endlessly. (グミ、手の届くところがあると無限に食べ続けてしまう。)	10	I also love gummies, especially the ones with lots of fruit juice. (グミ私も好きです。特に果汁多めのやつが)	AN	1.0	1.0	0.0	1.0	0.75
Does anyone want to play Apex Legends ranked match with me sometime? (エペランクマ今度一緒にやってくれる人いませんか?..)	17	Maybe we can play a game together tonight. (せっかくだから今晚やってもいいですよ)	TN	1.0	1.0	1.0	1.0	1.00
Oh! I just realized I have over 1000 followers! (え!! 今気づいたんだけどふおろわさん1000人いてる!?)	10	Congrats! I'm glad to hear the good news before I go to bed. (おめでとう! 寝る前に良いニュース聞けて嬉しいよ。)	TU	1.0	0.0	1.0	0.5	0.63
Maybe I'll eat curry today. (今日はカレー食うかな)	17	Autumn goes by so fast. (秋はあっという間に過ぎていくよね)	AU	0.0	1.0	0.0	0.5	0.38

Table 2: Four example tuples $\langle u, t, r, c \rangle$ and their obtained SU, ST, S, I, and SSI-t values. All utterances and responses were originally written in Japanese and translated into English by the authors.

Category	SU	ST	S	I	SSI-t
AN	0.98	0.86	0.14	0.66	0.66
TN	0.97	1.00	0.80	0.76	0.89
TU	0.93	0.19	0.88	0.71	0.68
AU	0.06	0.87	0.13	0.66	0.43

Table 3: SU, ST, S, I, and SSI-t scores of responses of each NUTA category

time t , which evaluates whether r is specific to spoken time t . Consequently, r has a high $\text{TD}@t$ if it is natural at t but unnatural at another time and a low TD (i.e., near zero) if its naturalness remains unchanged as time changes. On the other hand, $\text{TD}(r | u)$, denoted by $\text{TD}@all$, is the CN from the most unnatural time t^U to the most natural time t^N , which evaluates whether r is specific to time or not. So, r has a high $\text{TD}@all$ if it is natural at some time but not at another time. Since we consider the naturalness of **T** responses of the NUTA categories varies with time, r is classified as **T** (resp. **A**) iff $\text{TD}(r | u)$ is high (resp. low).

NUTA category By the definitions of TAN $p(r | u, t)$ and $\text{TD}@all$ $\text{TD}(r | u)$, we define each NUTA category by Table 4. For instance, response r is classified as **TN** iff both $p(r | u, t)$ and $\text{TD}(r | u)$ are high, and r is classified as **AU** iff both $p(r | u, t)$ and $\text{TD}(r | u)$ are low. Strictly speaking, to use this definition, two thresholds must be set that distinguish between the high and low of $p(r | u, t)$ and $\text{TD}(r | u)$; however, since we believe that determining these thresholds in advance is difficult, all the methods proposed in this paper are designed so that they do *not* require such thresholds.

Category	$p(r u, t)$	$\text{TD}(r u)$
AN	High	Low
TN	High	High
TU	Low	High
AU	Low	Low

Table 4: Definition of NUTA categories using TAN $p(r | u, t)$ and $\text{TD}@all$ $\text{TD}(r | u)$.

Related criteria In natural language processing, various criteria have been proposed for measuring the dependency between two sentences. Li et al. (2016) proposed *pointwise mutual information* (PMI) to choose appropriate response r to given utterance u :

$$\text{PMI}(r | u) \triangleq \log \frac{p(r | u)}{p(r)}. \quad (7)$$

As an extension of PMI, Paranjape and Manning (2021) proposed *pointwise conditional mutual information* (PCMI) to cope with additional external information other than utterance u to evaluate an appropriate response to u . Given utterance u , response r , and time information t as external information, $\text{PCMI}@t$ is defined as

$$\text{PCMI}(r | u, t) \triangleq \log \frac{p(r | u, t)}{p(r | u)}, \quad (8)$$

where let $p(r | u)$ be the *time-unaware naturalness* (TUN). While $\text{TD}@t$ in Eq. (5) is the CN of r given u from the most unnatural time t^U to the current time t , $\text{PCMI}@t$ is the CN of r given u when the naturalness changes from TUN to TAN at t . In a similar manner as $\text{TD}@all$, we define $\text{PCMI}@all$ as $\text{PCMI}(r | u) \triangleq \max_{t \in \mathcal{T}} \text{PCMI}(r | u, t)$.

We believe our TD is a more appropriate metric for time information than PCMI. PCMI considers the presence or absence of time information, not its change; however, time information always exists and changes, unlike such common external information as user persona and knowledge graphs. Therefore, PCMI is expected to be more blurred in its evaluation than TD. For instance, suppose two time ranges, t_1 and t_2 , such that $p(t_1) = p(t_2) = 0.5$ where (u, r) is a strongly time-specific response such that $p(r | u, t_1) = 0$ (i.e., r cannot be a response to u at t_1) and $p(r | u, t_2) = 1$ (i.e., r is a perfect response to u at t_2). Then, $p(r | u) = 0.5$ since $p(r | u) = \sum_{t \in \{t_1, t_2\}} p(t)p(r | u, t)$. By Eqs. (3), (5) and (6), $\text{TD}(r | u) = \infty$ where TD@all considers r as a strongly time-specific response. On the other hand, $\text{PCMI}(r | u, t) = \ln 2 \approx 0.69$ where PCMI@all considers r as not so time-specific. Thus, using TUN $p(r | u)$ blurs the evaluation of time dependency, and our TD@all is expected to detect time-specific responses more clearly than PCMI@all.

4.2 Experiments

We conducted experiments to show that our automatic metric for the NUTA categories correctly orders responses of the NUTA dataset of Section 3.1.

4.2.1 Experimental Settings

For each utterance $u \in \mathcal{U}$, we ranked four tuples containing u in the NUTA dataset by each quantity: TUN $p(r | u)$, TAN $p(r | u, t)$, TD@all $\text{TD}(r | u)$, TD@t $\text{TD}(r | u, t)$, PCMI@all $\text{PCMI}(r | u)$, and PCMI@t $\text{PCMI}(r | u, t)$. We attached a label high (resp. low) to the top (resp. bottom) of two tuples in each ranking for each quantity and used the obtained labels to classify NUTA categories.

As TUN $p(r | u)$, we used the Transformer-based Japanese dialogue model (TJD) (Sugiyama et al., 2023) with 1.6B parameters trained on more than two billion tweet-reply pairs: (u, r) . We constructed TAN $p(r | u, t)$ by fine-tuning the above model using tweet-reply-time triplets: (u, r, t) , where the fine-tuning dataset was obtained similarly as (Sato et al., 2017): we collected Japanese tweets with replies from August 2021 to April 2022 with filtering rules described in Appendix A.1 and obtained 470,255,625 triplets. We denote the fine-tuned time-aware TJD by TJD-t and used TJD $p(r | u)$ and TJD-t $p(r | u, t)$ to evaluate TD@all/t and PCMI@all/t. Detailed implementational settings are shown in Appendix A.2.

	Naturalness		TD		PCMI	
	TUN	TAN	@all	@t	@all	@t
N	0.60	0.68	0.47	0.58	0.46	0.55
T	0.54	0.54	0.68	0.60	0.53	0.48

Table 5: Accuracy of N/U and T/A classifications of each quantity. Best scores are indicated by **bold**.

	TD		PCMI	
	@all	@t	@all	@t
AN	0.54	0.43	0.43	0.38
TN	0.48	0.60	0.35	0.51
TU	0.44	0.19	0.34	0.14
AU	0.45	0.44	0.33	0.32
Ave.	0.48	0.42	0.36	0.34

Table 6: Accuracy of each NUTA category of each combination. Best scores are indicated by **bold**.

4.2.2 Experimental Results

We first checked whether TAN $p(r | u, t)$ and TD@all $\text{TD}(r | u)$ correctly classified **N** and **T**. Using high/low labels obtained by each quantity, we categorized tuples with high (resp. low) labels as **N** (resp. **U**) and computed the accuracy of the **N** category of each quantity. Similarly, we also computed the accuracy of the **T** category. Table 5 shows the accuracies of each quantity and indicates that TAN and TD@all achieved the highest accuracy of **N** and **T**. Consequently, TAN and TD@all are appropriate quantities for evaluating the naturalness and time dependency of responses.

Next, we checked whether the combination of TAN and TD@all correctly classified each NUTA category. We categorized tuples into one of **AN**, **TN**, **TU**, and **AU** according to Table 4 using the combination of high/low labels obtained by TAN and those obtained by one of TD@all, TD@t, PCMI@all, and PCMI@t. Table 6 shows the accuracy of each NUTA category of each combination and indicates that TD@all achieved the best average accuracy; however, for the **TN** category, TD@t achieved the best. Since TD@t evaluates whether response r is specific to the current time t but not to other times, it is more effective to detect **TN** responses than **TU** responses. Because our original motivation for using these quantities as automatic metrics for the NUTA categories is to detect **TN** responses that users prefer than **AN**, we conclude that TD@t $\text{TD}(r | u, t)$ is the most appropriate metric for constructing a time-considerable dialogue model.

Model	len	distinct-1	distinct-2	SU	ST	S	I	SSI-t
TJD	17.22	0.36	0.57	0.86	0.84	0.16	0.40	0.56
TC-TJD	12.50	0.55	0.78	0.86	0.85	0.19	0.45	0.59
TJD-t	16.74	0.33	0.48	0.78	0.83	0.12	0.36	0.52
TC-TJD-t	14.44	0.49	0.75	0.86	0.86	0.20	0.48	0.60
GPT-3.5	40.50	0.45	0.76	0.80	0.89	0.36	0.60	0.66
TC-GPT-3.5	34.92	0.44	0.74	0.74	0.90	0.53	0.54	0.68
GPT-4	37.79	0.40	0.73	0.86	0.96	0.89	0.62	0.83
TC-GPT-4	30.88	0.37	0.69	0.83	0.96	0.93	0.59	0.83

Table 7: Average length (len), distinct- N ($N = 1, 2$), SU, ST, S, I, and SSI-t scores of each model. TC- \mathcal{M} is a time-considerable \mathcal{M} achieved by our proposed reranking method. White and gray rows correspond to \mathcal{M} and TC- \mathcal{M} . If TC- \mathcal{M} 's score exceeded \mathcal{M} ' one, it is indicated by **bold**.

5 Time-Considerable Dialogue Models

We propose a new reranking method to make existing dialogue models *time-considerable*. We applied our method to various existing models, including GPT-4, which is a state-of-the-art LLM, and evaluated the time-considerable dialogue models to verify whether they improved response qualities.

5.1 Proposed Reranking method

Let \mathcal{M} be any dialogue model that can generate multiple responses to the same utterance. Our proposed reranking method extends \mathcal{M} to be time-considerable. Given base dialogue model \mathcal{M} , TAN $p(r | u, t)$, positive integer N , and probability δ , we obtain time-considerable response r^* to utterance u at time t by the following manner:

1. Generate N candidate responses to u at t , denoted by $\mathcal{R} \triangleq \{r_i | i \in [N]\}$, from base dialogue model \mathcal{M} ,
2. Evaluate TAN $p(r_i | u, t)$ for all $r_i \in \mathcal{R}$ and delete r_i from \mathcal{R} if r has no sufficient naturalness; i.e., $p(r_i | u, t) \leq \delta$,
3. Evaluate TD@ t $\text{TD}(r_i | u, t)$ for all $r_i \in \mathcal{R}$ and find the most time-specific response $r^* \in \max_{r \in \mathcal{R}} \text{TD}(r | u, t)$,
4. Return obtained r^* as a time-considerable response to u at t .

Since Step 2 removes candidate responses with lower naturalness than threshold δ , the filtering mechanism may improve the naturalness of the final response r^* when the response generation model is weak.

Our proposed reranking method is simple but strong because we can create various time-considerable dialogue models by combining existing base dialogue models and TANs, where

base dialogue model \mathcal{M} is required only to generate multiple responses for the same utterance and TAN $p(r | u, t)$ only to be evaluable. Namely, our method can be applied to dialogue models whose architectures and parameters are not publicly available but are provided as APIs. Of course, if \mathcal{M} is time-aware and evaluable, it can also be used as TAN $p(r | u, t)$.

5.2 Experiments

We applied our proposed reranking method to existing dialogue models and gauged their performance by human subjective evaluations.

5.2.1 Experimental Settings

We briefly explain our experimental settings, and the detailed settings are shown in Appendix A.2.

Base dialogue models We used the following four dialogue models as base dialogue model \mathcal{M} of our proposed reranking method:

1. **TJD** is a transformer-based Japanese dialogue model with 1.6B parameters trained on over two billion tweet-reply pairs (Sugiyama et al., 2023) described in Section 4.2.
2. **TJD-t** is a time-aware TJD obtained by fine-tuning described in Section 4.2.
3. **GPT-3.5** is an extension LLM of GPT-3 (Brown et al., 2020) with 355B parameters and supports various tasks in many languages. gpt-3.5-turbo is a specialized GPT-3.5 for dialogue tasks and is provided as an API.
4. **GPT-4** is a large-scale multimodal model that extends GPT-3.5 (OpenAI, 2023). gpt-4 is a specialized GPT-4 for dialogue tasks and is provided as an API.

All models can generate multiple responses for the same utterance by top- p sampling (Holtzman et al., 2020), where p is a hyperparameter and set to 0.9 through the experiments. Since GPT-3.5 and GPT-4 do not treat time information as input, they are originally time-unaware; however, in our experiment, we used them with prompts to generate time-aware responses. We created prompts based on a sample prompt for dialogue tasks provided by OpenAI and shown in Appendix A.2.

Settings on proposed reranking method For any base dialogue model \mathcal{M} , we denote time-considerable \mathcal{M} achieved by our proposed reranking method by TC- \mathcal{M} (e.g., TC-TJD, TC-GPT-4). Throughout the experiments, we used TJD-t as TAN $p(r | u, t)$ and set $N = 20$ and $\delta = 0$, where N is the number of candidate responses to be generated and δ is a threshold for filtering candidate responses by their naturalness. We set $\delta = 0$ because we used sufficiently strong models for response generation and did not aim to improve their naturalness.

Time intervals Since TD@ t is the maximum of $\text{CN}_{t^U:t}(r | u)$ for all possible $t^U \in \mathcal{T}$, t^U could be very close to spoken time t , and in such cases, the time dependency of r might be not interpretable for humans since it is too *short-term*. To avoid detecting such short-term time dependency, we divided 24 hours into three intervals, morning (3 to 8), noon (9 to 17), and night (18 to 2), and defined their representatives as 6, 13, and 22, with reference to Yamamoto and Shimada (2019). The reason of adopting the above intervals is that morning/noon/night is defined based on the human life-cycle; thus, this interval is more intuitive for human understanding. We evaluated TD@ t by Eq. (5) with \mathcal{T} as representatives except for one of spoken time t (e.g., $t = 10 \Rightarrow \mathcal{T} = \{6, 22\}$).

Evaluation dataset As an evaluation dataset, we prepared 100 utterance-time pairs in the same manner as \mathcal{U} in Section 3.1, where we excluded the same tweets as the NUTA dataset. Given utterance u and its spoken time t , for each base dialogue model $\mathcal{M} \in \{\text{TJD}, \text{TJD-t}, \text{GPT-3.5}, \text{GPT-4}\}$, we generated the best response of \mathcal{M} and a time-considerable response of TC- \mathcal{M} , denoted by $\bar{r}_{\mathcal{M}}$ and $r_{\mathcal{M}}^*$. Thus, the evaluation dataset consisted of 100 tuples of u, t , and responses $\bar{r}_{\mathcal{M}}$ and $r_{\mathcal{M}}^*$ for each \mathcal{M} .

Evaluation criteria For each dialogue model, we computed the average output response length (len) and distinct- N ($N = 1, 2$), which is the ratio of the number of unique N -grams to the total words generated by the dialogue model, for evaluating its diversity (Li et al., 2016). We also obtained the ST, SU, S, I, and SSI-t scores defined in Section 3.2 by human subjective evaluations. We assigned five crowd workers for each response and asked them to evaluate its SU, ST, S, and I scores with a five-point Likert scale (1 to 5) and normalized them into $[0, 1]$ scale. We obtained the SU, ST, S, and I values of each model by averaging the annotated values of responses generated by the model. Finally, we obtained its SSI-t value as the average of those values.

5.2.2 Experimental Results

Table 7 shows the values of each model for the eight indicators: the average output response length (len), distinct- N ($N = 1, 2$), SU, ST, S, I, and SSI-t scores. For any score except len, a higher value denotes a better result. In the table, if the score of TC- \mathcal{M} exceeds the original \mathcal{M} , we highlighted it in **bold**; in other words, bold scores were improved by our proposed reranking method.

TC- \mathcal{M} improved the ST, S, and SSI-t scores for almost all \mathcal{M} ; namely, it was more time-considerable than the original \mathcal{M} . TC- \mathcal{M} tended to prefer a short response to \mathcal{M} because the len of TC- \mathcal{M} is smaller than \mathcal{M} for each model. For GPTs, the distinct- N , SU, and I scores of TC- \mathcal{M} were slightly lower than \mathcal{M} . This is because those scores of \mathcal{M} were already sufficiently high, and our reranking method aims to improve time-considerability but not diversity and general sensibility of naturalness. Consequently, we confirmed that our reranking method improved the quality of the existing dialogue models.

One interesting observation from this experiment is that TJD-t, which we used as TAN $p(r | u, t)$, had lower quality than TJD. This fact suggests that fine-tuning explained in 4.2 was not enough and that our reranking method improved the performance of the existing dialogue models even if TAN’s performance is somewhat low. However, of course, using TAN with low performance is not desirable. For GPTs, TC- \mathcal{M} achieved a higher S score but lower SU and I scores than the original \mathcal{M} . We believe this deterioration of SU and I scores can be easily avoided using a more accurate model as TAN, such as GPT-4.

Utterance	Tomo-san, you can get up feeling refreshed after drinking; I really like you. (tomoさんのお酒飲んだ次の日寝起きいいのすき) @ 10 a.m (daytime)
TJD	Tomo-san! Tomo-san! Tomo-san! Tomo-san! (とーもーさん!!とーもーさん!!とーもーさん!!とーもーさん!!)
TC-TJD	I overslept and missed my first class. (1限寝過ごしたよ)
TJD-t	I also get up feeling refreshed after drinking. (わたしもお酒飲んだ次の日寝起きいいの)
TC-TJD-t	I'm sure that I'll get back to bed. (二度寝安定ですねこれは)
GPT-3.5	Yes, I have trouble getting up after drinking, but I slept well last night, so I feel fine. (そうですね、お酒を飲んだ翌日は寝起きが悪くなりますが、今日はしっかりと寝ていたので大丈夫です。)
TC-GPT-3.5	Yes, you feel better. But I was drinking last night, so I am a little hungover today. (そうですね、体が軽く感じますよね。でも、昨日は遅くまで飲んでしまったので、今日は少し眠いです。)
GPT-4	Oh, that's great! Being energetic in the morning has a very positive effect on the start of your day, doesn't it? (えー、それはすごいですね!朝から元気なのは、一日の始まりにとっても良い影響がありますよね。)
TC-GPT-4	Oh, that's great! Being energetic in the morning has a very positive effect on the start of your day, doesn't it? (えー、それはすごいですね!朝から元気なのは、一日の始まりにとっても良い影響がありますよね。)
Utterance	I still love a big chorus!!!! (やっぱり大合唱は最高だ!!!!) @ 7 p.m (nighttime)
TJD	Thanks a lot! (おつかれさまでした!)
TC-TJD	Yay! Thanks for today! It was fun! (いえーい!今日はありがとう!楽しかった!)
TJD-t	Thanks a lot! I had the best time!!!! (おつかれさまでした!最高にたのしかったです!!!)
TC-TJD-t	It was great...! Thank you so much for today! (最高だったね...!今日は本当にありがとう!)
GPT-3.5	That's right! That song is a real crowd-pleaser when everyone sings it together. But what about the next song? (そうだね!特にあの曲はみんなで歌うと本当に盛り上がるよね。でも、次の曲はどうする?)
TC-GPT-3.5	Yes, a big chorus is really exciting, isn't it? Are you going out for a drink afterward? (そうだね、大合唱は本当に盛り上がるよね。この後は何か飲みに行く予定はある?)
GPT-4	I really agree! It's even more exciting to hear a big chorus at this time of night , isn't it? (本当にそうだね!夜のこの時間に大合唱を聴くと、さらに感動が増すよね。)
TC-GPT-4	I really do! But it's almost time for dinner , so why don't we head home? (本当にそうだね!でもそろそろ夕食の時間も近いから、帰ろうか?)

Table 8: Examples of generated responses by each model. TC- \mathcal{M} is time-considerable \mathcal{M} achieved by our proposed reranking method. White and gray rows correspond to \mathcal{M} and TC- \mathcal{M} . Time-specific phrases are indicated by **bold**. All utterances and responses were originally written in Japanese and translated into English by the authors.

Table 8 shows example responses $\bar{r}_{\mathcal{M}}$ and $r_{\mathcal{M}}^*$ generated by each \mathcal{M} and TC- \mathcal{M} , where $\bar{r}_{\mathcal{M}}$ and $r_{\mathcal{M}}^*$ are shown in white and gray rows. The result indicates $r_{\mathcal{M}}^*$ contained more time-specific expressions than $\bar{r}_{\mathcal{M}}$; e.g., “I overslept” in the morning and “thanks for today” at night. The detailed analysis is described in Appendix A.3. For the first utterance, the original GPT-4 already generated time-specific responses, and TC-GPT-4 selected the same response as its output; namely, in this example, TC-GPT-4 considered that the original response was sufficiently time-considerable.

6 Conclusion

We proposed a new reranking method to construct time-considerable dialogue models that distinguish between always natural responses AN and temporally natural responses TN and actively output TN. We verified the assumption *users prefer TN to AN* by human evaluations and introduced a new metric to classify the NUTA categories of responses. We proposed a new reranking method to make existing dialogue models *time-considerable* using the metric and empirically showed that our

method improved the qualities of existing models.

A promising future study is to control the degree to which a time-considerable dialogue model actively uses time information in different situations. Not only time information but all external information is valuable when used appropriately, but excessive use may harm the users' impression. Therefore, we plan to develop a mechanism to estimate an appropriate TD@ t value that a response should have in the current situation and to output or generate a response with the estimated value.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19H05693.

Limitations

Effect of TAN's quality In our proposed reranking method, we used TJD-t, which is a time-aware transformer-based Japanese dialogue model obtained by fine-tuning, as TAN, but our experimental result showed the quality of TJD-t was lower than the other models. One of our contributions is that we empirically showed that our reranking method

successfully improved the S (specificity to time) score of each model even though the quality of TAN (TJD-t) is somewhat low. However, it would be desirable to investigate how the effectiveness of our proposed method changes as the quality of TAN changes.

Different types of time information In this paper, we used the time of day as time information. However, there are other types of time information that have longer periods, such as day of the week and seasons. It is promising future research to investigate how the quality of dialogues changes with the use of such longer periodic time information.

Cultural differences in time information In this paper, we investigate the effect of the use of time information on dialogues in Japanese; however, it has been shown that the time-specific expression varies depending on the country and culture (Shwartz, 2022). Therefore, it is desirable to investigate whether the proposed method can produce time-considerable responses for different languages.

Ethics Statement

In this paper, we employed workers using a crowdsourcing service. We made sure that the workers were paid above the minimum wage. It applies to all crowdsourcing experiments in this paper.

References

- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. <https://doi.org/10.48550/arXiv.2001.09977>.
- Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoun Kang, Sang-Woo Lee, and Woomyoung Park. 2022. Building a role specified open-domain dialogue system leveraging large-scale language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2128–2150.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems (NeurIPS)*, 33:1877–1901.
- Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5114–5132.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. <https://arxiv.org/abs/1904.09751>.
- Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2022. Building a personalized dialogue system with prompt-tuning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) Student Research Workshop*, pages 96–105.
- Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71.
- Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating personalized dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3045–3059.
- 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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 110–119.
- Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Carl Yang, and Liang Zhao. 2023. Beyond one-model-fits-all: A survey of domain specialization for large language models. <https://arxiv.org/abs/2305.18703>.

- Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhunoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL*, pages 1317–1337.
- Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022. Partner personas generation for dialogue response generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5200–5212.
- Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.
- Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2022. Achieving conversational goals with unsupervised post-hoc knowledge injection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3140–3153.
- Teruhisa Misu. 2018. Situated reference resolution using visual saliency and crowdsourcing-based priors for a spoken dialog system within vehicles. *Computer Speech 'I&' Language*, 48:1–14.
- OpenAI. 2023. Gpt-4 technical report. <https://arxiv.org/abs/2303.08774>.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744.
- Ashwin Paranjape and Christopher Manning. 2021. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 768–781.
- Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 690–695.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–325.
- Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2017. Modeling situations in neural chat bots. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) Student Research Workshop*, pages 120–127.
- Iulian Serban, Alessandro Sordani, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 30.
- Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL*, pages 2842–2853.
- Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based japanese chat systems. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, pages 685–691.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lambda: Language models for dialog applications. <https://doi.org/10.48550/arXiv.2201.08239>.
- Dina Utami and Timothy Bickmore. 2019. Collaborative user responses in multiparty interaction with a couples counselor robot. In *Proceedings of ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 294–303.
- Oriol Vinyals and Quoc Le. 2015. A neural conversational model. <https://arxiv.org/abs/1506.05869>.
- Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving knowledge-aware dialogue generation via knowledge base question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34:05, pages 9169–9176.
- Kohei Yamamoto and Kazutaka Shimada. 2019. Acquisition of knowledge with time information from twitter. In *Proceedings of International Conference on Asian Language Processing (IALP)*, pages 148–153.
- Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2015. Recurrent neural network regularization. <https://doi.org/10.48550/arXiv.1409.2329>.

- Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2031–2043.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213.
- Wangchunshu Zhou, Qifei Li, and Chenle Li. 2021. Learning from perturbations: Diverse and informative dialogue generation with inverse adversarial training. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) and the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 694–703.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. <https://arxiv.org/abs/1909.08593>.

A Appendix

A.1 Dataset Construction

We describe the filtering rules used to obtain the NUTA dataset in Section 3 and the fine-tuning dataset in Section 4.

NUTA dataset In Section 3, we constructed the NUTA dataset which is a collection of tuples of utterances, their spoken times, their responses, and their NUTA categories. As utterances, we extracted Japanese tweets posted between May and December 2022 that satisfied the following conditions:

- Do not contain URLs, usernames, other tweets, parentheses,
- Consists of 6 to 30 characters,
- Not posted by users whose names contain “bot”,
- Not Replied to another tweet.

We randomly selected 1,000 tweets from the extracted tweets and manually deleted tweets containing discriminatory, violent, or inappropriate expressions. As a result, we obtained 640 appropriate tweets and used them as utterances \mathcal{U} .

Fine-tuning dataset In Section 4, we constructed a fine-tuning dataset in the same manner as [Shwartz \(2022\)](#) to obtain time-aware TJD by training TJD on the obtained dataset. We collected Japanese tweet-reply pairs with their timestamp posted between August 2021 to April 2022 that do not contains URLs or other tweets. As a result, we obtained 470,255,625 triplets.

A.2 Implementational and Experimental Settings

We here describe the detailed settings of our implementation and experiments.

TJD and TJD-t For a time-unaware dialogue model, we used TJD which is a transformer-based Japanese dialogue model with 1.6B parameters trained on over two billion tweet-reply pairs ([Sugiyama et al., 2023](#)). We downloaded the trained TJD ³ and obtained a time-aware TJD, denoted by TJD-t, by fine-tuning TJD on Fairseq ⁴ ([Ott et al., 2019](#)), which is a sequence modeling toolkit to train custom models for various takes including translation, summarization,

³[https://github.com/nttcs-lab-japanese-dialog-transformers](https://github.com/nttcs/nttcs-lab-japanese-dialog-transformers)

⁴<https://github.com/facebookresearch/fairseq>

language modeling, and other text generation tasks. In fine-tuning, we used SentencePiece ⁵ ([Kudo and Richardson, 2018](#)) to tokenize utterances and responses written in Japanese. Table 9 shows the hyperparameters we set to in fine-tuning. We used the computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by the National Institute of Advanced Industrial Science and Technology (AIST).

Configurations	Values
Model Architecture	Transformer
Pretrained Model	TJD (Sugiyama et al., 2023)
Devices	Nvidia V100 GPU
Max tokens	4,000
Optimizer	Adafactor
Learning rate	1e-04
Learning rate scheduler	inverse sqrt
Warmup	10000
weight decay	0.0
Loss Function	label smoothed cross entropy

Table 9: Hyper-parameters for fine-tuning

GPT-3.5/4 We used GPT-3.5 and GPT-4 ([OpenAI, 2023](#)) as state-of-the-art dialogue models. Since architectures and parameters of GPT-3.5/4 were not publicly available, we used OpenAI APIs `gpt-3.5-turbo` and `gpt-4` to generate responses of GPT-3.5/4. We created the following prompt based on a sample prompt for dialogue tasks provided by OpenAI and gave the prompt as input to GPT-3.5/4 to generate a time-aware response.

The current time is [hour], and A and B are having a conversation. Taking into account the current time, generate the following B's response to A's utterance. However, avoid expressions like "it's [hour] o'clock now."

A: [utterance]

B:

TD@ t in our proposed reranking method As shown in Eq. (5), $TD@t$ is the maximum of $CN_{t^U,t}(r|u)$ for all possible $t^U \in \mathcal{T}$, and t^U could be very close to spoken time t . In such cases, the time dependency of r might be difficult for humans to understand since its naturalness varies in too *short-term*. To avoid detecting such short-term time dependency, we divided 24 hours into three intervals, morning (3 to 8), noon (9 to 17),

⁵<https://github.com/google/sentencepiece>

and night (18 to 2). We used 6, 13, and 22 as the representatives of morning, noon, and night, respectively. The division and their representatives were determined with reference to Yamamoto and Shimada (2019). We computed TD@ t by Eq. 5 using the above \mathcal{T} excluding the representative of spoken time t . For instance, for $t = 10$, we use $\mathcal{T} = \{6, 22\}$ to evaluate TD@ t because the representative of 10 is defined as 13. We conducted the same experiments as Section 4 using the above TD@ t , and Table 10 and 11 show the results. The results indicate that the accuracy of TN classification was slightly improved by introducing the above time intervals but not for the average accuracy. Therefore, it cannot be said that either TD with intervals is better than the original TD.

	Naturalness		TD		PCMI	
	TUN	TAN	@all	@ t	@all	@ t
N	0.60	0.68	0.48	0.61	0.48	0.55
T	0.54	0.54	0.65	0.58	0.51	0.48

Table 10: Accuracy of N/U and T/N classifications with the morning/day/naight quantization

	TD		PCMI	
	@all	@ t	@all	@ t
AN	0.50	0.39	0.41	0.38
TN	0.48	0.61	0.39	0.51
TU	0.41	0.15	0.30	0.14
AU	0.46	0.45	0.34	0.32
Ave.	0.46	0.40	0.36	0.34

Table 11: Accuracy of NUTA classification with the morning/day/naight quantization

A.3 Time-specific Expressions

We conducted a subjective evaluation to count the number of time-specific expressions by authors. As a result, our method increased their occurrences by 80% and 8% for TJD-t and GPT-4, respectively. These results on occurrences of time-specific expression are consistent with the improvements in S shown in Table 7.