Dialogue Response Generation Using Completion of Omitted Predicate Arguments Based on Zero Anaphora Resolution

Ayaka Ueyama and **Yoshinobu Kano**
Shizuoka University
aueyama@kanolab.net, kano@inf.shizuoka.ac.jp

Abstract

Human conversation attempts to build *common ground* consisting of shared beliefs, knowledge, and perceptions that form the premise for understanding utterances. Recent deep learning–based dialogue systems use human dialogue data to train a mapping from a dialogue history to responses, but common ground not directly expressed in words makes it difficult to generate coherent responses by learning statistical patterns alone. We propose Dialogue Completion using Zero Anaphora Resolution (DCZAR), a framework that explicitly completes omitted information in the dialogue history and generates responses from the completed dialogue history. In this study, we conducted automatic and human evaluations by applying several pretraining methods and datasets in Japanese in various combinations. Experimental results show that the DCZAR framework contributes to the generation of more coherent and engaging responses.

1 Introduction

Dialogue systems for natural language conversation, dialogue, and discourse with humans have attracted widespread attention in industry and academia. Especially in recent years, the development of deep learning techniques and large dialogue corpus have made remarkable progress in dialogue response generation (Komeili et al., 2022; Borgeaud et al., 2022; Thoppilan et al., 2022). However, the performance of the dialogue systems is still unsatisfactory, and many problems remain to be resolved. One problem is that dialogue systems cannot accurately interpret the intent of human utterances because the construction of common ground, which is important in human-to-human dialogue, has not yet been established (Stalnaker, 1978; Clark and Schaefer, 1989). Common ground in dialogue refers to shared beliefs, knowledge, and perceptions that form the premise for understanding utterances. For example, much information is omitted in the dialogue in Table 1, but the two speakers can convey their intentions in short utterances because, through their common knowledge and context, they can omit information but still *understand each other*.

Speaker A:	My friend has not come to school.
	I'm worried ϕ_{DAT} [about my friend].
	Should I try to call ϕ_{DAT} [my friend]?
Speaker B:	Something could be wrong
	ϕ_{DAT} [for your friend].
	Perhaps ϕ_{NOM} [you should] try to
	call ϕ_{DAT} [your friend].

Table 1: **Example of dialogue where omission occurs.** Highlighted text represents omitted arguments.

Why has the construction of common ground not been realized in human-to-system dialogues? Sequence-to-Sequence (Seq2Seq) models (Sutskever et al., 2014; Cho et al., 2014) have been widely used in recent dialogue systems (Vaswani et al., 2017; Raffel et al., 2020; Lewis et al., 2020; Bao et al., 2020; Zhang et al., 2020b). Seq2Seq models use large amounts of dialogue data to train a mapping from a dialogue history to responses. However, there are many omissions in dialogue data, and it is difficult for models to generate responses that convey human intentions simply by training statistical patterns. To address this problem, several methods that use a knowledge base (KB) have been proposed. These models bridge the gap between humans and models by introducing external knowledge and providing the models with common-sense knowledge (Zhao et al., 2020; Eric et al., 2021; Xu et al., 2022). Human common-sense knowledge is one piece of information that can be omitted, but the cost of building a KB is significant and not easily transferable to different domains or models.

In this study, we considered a method to provide

models with omitted information without using external knowledge. Dialogue systems can precisely interpret the intent of human utterances only when the roles of involved persons and things are understood, but omissions frequently occur in Japanese dialogue to avoid repetition and references to self-evident objects (Seki et al., 2002). Thus, the coherence of responses can be improved by inferring and explicitly incorporating the roles of persons and things. Inspired by the idea of zero anaphora resolution (ZAR), we propose Dialogue Completion using Zero Anaphora Resolution (DCZAR), a framework that explicitly completes omitted information in a dialogue history and generates responses from the completed history.

The DCZAR framework consists of three models: a predicate-argument structure analysis (PAS) model, a dialogue completion (DC) model, and a response generation (RG) model. The PAS model analyzes the omitted arguments (*zero pronouns*) in the dialogue, and the DC model determines which arguments to complete and where to complete them and explicitly completes the omissions in the dialogue history. The RG model, trained by the complementary dialogue history and response pairs, generates a response. The PAS and RG models are constructed by fine-tuning the common pretrained model with a dataset corresponding to each task, while the DC model uses a pretrained model without fine-tuning. We used the Japanese Wikipedia dataset and Japanese postings ("tweets") to Twitter to build our pretrained models. Since tweets are like dialogues in that they contain many abbreviations and short sentences, the model pretrained with tweets is expected to improve the performance of ZAR and dialogue response generation.

In this study, we performed automatic and human evaluations of three models built by pretraining models constructed by combining different methods and datasets. Experimental results show that the DCZAR framework can be used to generate more coherent and engaging responses. Analysis of the responses shows that the model generated responses that were highly relevant to the dialogue history in dialogues with many characters. The three main contributions of this work are as follows:

- We show that incorporating argument omission completion based on ZAR into the RG model significantly improves the coherence and engagement of the responses (Section 4.5).
- ZAR performance is improved by pretraining with Twitter data that have similar features to the dialogue data (Section 4.3).
- We confirm that the DC model can complete dialogue omissions with sufficient performance (Section 4.4).

2 Related Work

2.1 Dialogue Response Generation

Dialogue response generation is the task of generating an appropriate response following a given dialogue history, and can be formulated as a serial transformation problem that generates a target sentence from a source sentence (Ritter et al., 2011; Serban et al., 2017; Zhou et al., 2022). Specifically, given a dialogue history $H = \{X_1, X_2, \ldots, X_n\}$ consisting of n turns (where $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,m}\}$ is an utterance consisting of m tokens), the problem is to approximate a model distribution that gives a generated response sentence $Y = \{y_1, y_2, \ldots, y_o\}$ consisting of the corresponding o tokens to the data distribution of the human response sentence $T = \{t_1, t_2, \ldots, t_p\}$.

$$P_\theta(Y \mid H) = \prod_{i=1}^{m} P_\theta\left(y_i \mid y_{<i}, X_1, \ldots, X_n\right) \quad (1)$$

2.2 Zero Anaphora Resolution

ZAR is the task of detecting any omitted arguments of a predicate and identifying its antecedents. It is formulated as part of the predicate-argument structure analysis task. In the NAIST Text Corpus (NTC) 1.5, the standard benchmark dataset for ZAR, each predicate is annotated with an argument representing either the nominative (NOM), accusative (ACC), or dative (DAT) case. A ZAR task is classified as *intra* (arguments in the same sentence in which the predicate appears), *inter* (arguments in a sentence preceding the predicate), or *exophora* (arguments not existing in the sentence), according to the positional relationship between a predicate and its arguments. If the argument of a predicate is directly dependent on the predicate, it is a syntactic-dependent argument (*dep*).

There has been extensive research on the application of ZAR to Japanese (Sasano and Kurohashi, 2011; Yamashiro et al., 2018; Umakoshi et al.,

2021). Konno et al. (2021) proposed a new pretraining task and a fine-tuning method for ZAR, assuming the importance of common-sense knowledge to understand the contextual connections around zero pronouns and antecedents.

Pseudo Zero Pronoun Resolution (PZERO). PZERO focuses on the acquisition of common-sense knowledge. It is a pretraining task that replaces one of the noun phrases that occur two or more times in the input series with a mask token (`[MASK]`) and selects from the input series the token that should be filled in for `[MASK]`. Since the task of selecting `[MASK]` from the input series is similar to the task of identifying the antecedent corresponding to a zero pronoun, we expect the model to acquire the common-sense knowledge required for ZAR. The model takes as input a series $X = \{x_1, x_2, \ldots, x_T\}$ of length T containing `[MASK]`, and selects a token from the series X at the end of the noun phrase that should be filled in for `[MASK]` as the result. All noun phrases that have the same letter as the masked noun phrase are considered correct.

Argument Selection as Pseudo Zero Pronoun Resolution (AS-PZERO). AS-PZERO is a method of parsing predicate arguments in the same format as PZERO, using parameters trained in PZERO. The model takes as input a series X and the predicates it contains, and selects from the input series the token with the highest likelihood as the result of guessing the word that is the argument of the predicate. If the predicate argument is not present in the input series X, let the model select `[CLS]`, and once `[CLS]` is selected, further classify arguments into four categories (author, reader, general, or none). The probability distribution for each category is obtained from the node which corresponds to the `[CLS]` token in the final layer.

3 Approach: DCZAR Framework

We propose the DCZAR framework, which, as mentioned in Section 1, consists of three models: PAS, DC, and RG. Figure 1 shows an overview of the proposed DCZAR framework.

3.1 PAS Model

The PAS model performs a predicate-argument structure analysis on the input dialogue history $X = \{x_1, x_2, \ldots, x_T\}$ of length T and predicts the arguments $A_{case} = \{a_{case,1}, a_{case,2}, \ldots, a_{case,n}\}$, where $case \in \{\texttt{NOM}, \texttt{ACC}, \texttt{DAT}\}$ and represents the case information, corresponding to the n predicates $P = \{p_1, p_2, \ldots, p_n\}$.

3.2 DC Model

Using the dialogue history X, the predicates P, and the arguments A_{case} predicted by the PAS model, the DC model explicitly complements omissions in the dialogue history to create multiple candidate sentences, calculates scores representing the sentence naturalness, re-ranks the sentences based on that score, and selects the sentence with the highest score. When complementing, it is necessary to determine *whether the argument should be completed* and *where it should be complemented*.

Word order is relatively flexible in Japanese, but a sentence becomes unnatural when argument types and their order is not relevant. The location of the argument completion is thus important. To determine whether an argument should be completed, first check whether there is an argument $a_{case,i}$ between a predicate p_i and the predicate p_{i-1} preceding it (search range r_i); if not, then $a_{case,i}$ is to be completed. Next, regarding where it should be complemented, pseudo-log-likelihood scores (PLLs) (Salazar et al., 2020), a measure of sentence naturalness, determines the position of completion. PLLs measure the sum of the log-likelihoods of the conditional probabilities of predicting the replacement of each token with `[MASK]`, with more natural sentences having higher scores. To determine the position of completion, the target token of completion is inserted between each token in the search range, multiple candidate sentences are created, and PLLs are calculated for all candidate sentences. For example, if there are n tokens to be completed and m tokens in the search range, the number of candidate sentences is expressed as

$$\sum_{k=0}^{n} \frac{{}_nC_k(m+n-k)!}{m!} \tag{2}$$

The sentence with the highest score is then selected and used as input for the RG model.

3.3 RG Model

The RG model is trained by the dialogue history and response pairs are selected by the DC model. Only the dialogue history is used as input for response generation during inference.

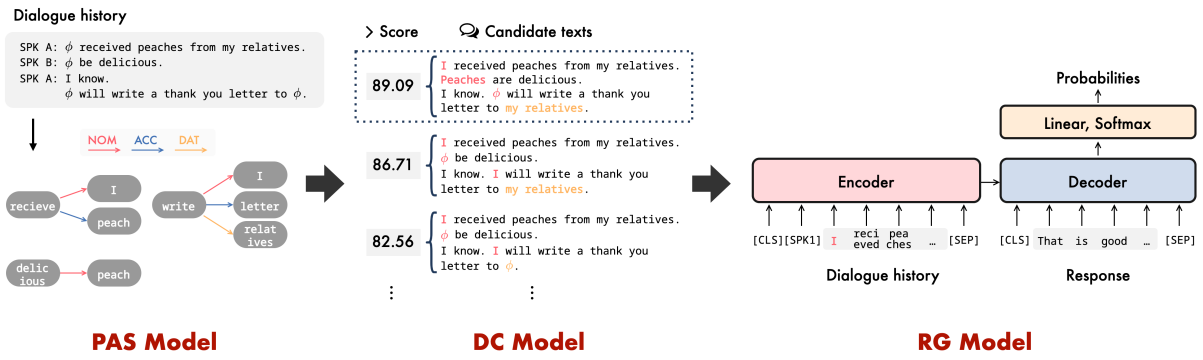

Figure 1: **Overview of our approach, the DCZAR framework.** The PAS model analyzes the omitted arguments (zero pronouns) in the dialogue history, the DC model determines which arguments to complete and where to complete them, and explicitly completes omissions in the dialogue history. The RG model, trained by the complementary dialogue history ($1, 2, \ldots, n-1$-th utterances) and response (n-th utterance) pairs, generates a response.

4 Experiments

4.1 Pretraining Setup

In this work, we constructed four pretraining models using two pretraining methods and two pretraining datasets in combination, and verified which model achieved better performance on each task. This section describes the construction of the pretrained models. The pretrained models described in this section are used in Section 4.3, 4.4, and 4.5.

4.1.1 Pretraining Task

Cloze (Devlin et al., 2019) and PZERO (Konno et al., 2021) are used as pretraining tasks. We use the pretrained parameters of the bert-base-japanese-whole-word-masking model as the initial parameters of the model.

Cloze. Cloze is a pretraining task for a masked language model (MLM) that performs operations (replacing 80% of tokens with `[MASK]` and 10% with a random vocabulary token, and performing no operation on the remaining 10%) on 15% of tokens randomly selected from the input series, excluding `[CLS]` and `[SEP]`, and predict the tokens replaced with `[MASK]`.

PZERO. PZERO is a pretraining task that replaces one of the noun phrases that occurs two or more times in the input series with `[MASK]` and selects from the input series the token that should be filled in for `[MASK]`.

4.1.2 Dataset

We used Japanese Wikipedia and Japanese tweets collected on Twitter as the pretraining dataset.

Wikipedia. The Wikipedia dataset is a preprocessed dataset from Japanese Wikipedia, consisting of a training set of 15M sentences (763M tokens) and a development set of 3K sentences (about 220K tokens). As preprocessing, we removed XML tags, article titles, and URLs contained in the articles. When using these data in PZERO, it is necessary to identify noun phrases. Therefore, we identified noun phrases based on the analysis results of the morphological analyzer MeCab and the dependency analyzer CaboCha (Kudo and Matsumoto, 2002), as in the method of Konno et al. (2021). We used the BertJapaneseTokenizer to segment the text into subword units.

Twitter. The Twitter dataset is a preprocessed dataset of tweets collected using the Twitter API, consisting of a training set of 70M sentences (504M tokens) and a development set of 30K sentences (about 200K tokens). We removed mentions (alphanumeric strings beginning with @), hashtags (strings beginning with #), URLs, and pictograms as preprocessing. For noun phrase identification and subword segmentation, we employed the same method as used for the Wikipedia data.

4.2 Compared Models

We compared the combinations of pretrained models shown in Table 2. As mentioned earlier, we used two datasets (Wikipedia and Twitter), and two tasks (Cloze and PZERO) for pretraining, resulting in four combination patterns. We compared these four combinations throughout the PAS, DC, and RG models; for example, RG`wiki-cloze` model uses PAS`wiki-cloze` model and DC`wiki-cloze` model as its preprocessing, corresponding to patterns (e) to

ID	PAS Model	DC Model	RG Model
(a)	N/A	N/A	wiki-cloze
(b)	N/A	N/A	twitter-cloze
(c)	N/A	N/A	wiki-pzero
(d)	N/A	N/A	twitter-pzero
(e)	wiki-cloze	wiki-cloze	wiki-cloze
(f)	twitter-cloze	twitter-cloze	twitter-cloze
(g)	wiki-pzero	wiki-cloze	wiki-pzero
(h)	twitter-pzero	twitter-cloze	twitter-pzero

Table 2: **Compared patterns of pretrained models used for the PAS, DC, and RG models.** Patterns (a) to (d) are baseline response generation models, and patterns (e) to (h) are proposed models applying the DCZAR framework.

(h) in Table 2, where pattern (h) is the final combination with proposed pretrained models (twitter-pzero) only. We prepared baseline models, patterns (a) to (d) in Table 2, which do not apply any completion. An exception is that we do not use the PZERO task but the Cloze task for the DC model because the PLLs used in the DC model's complementary location prediction require a pretrained model that can solve the Cloze task.

4.3 Experiment 1: PAS Model

We evaluated the performance of the predicate argument structure analysis of the PAS models within patterns (e) to (h) shown in Table 2. The PAS models were pre-trained models with fine-tuning by the AS-PZERO task using NTC. The input to the PAS model was a sentence containing the predicate and its antecedent, and the PAS model is trained to output the antecedent and case information corresponding to the predicate.

4.3.1 Dataset

We used the NTC (Iida et al., 2010) to fine-tune the PAS model. This corpus is annotated with information on predicate-argument structures and coreference. In this study, we divided data into training, development, and test sets, following the method described in Taira et al. (2008). The numbers of *intra*, *inter*, and *exophora* for the training, development, and test instances were respectively 14K/3K/6K, 9K/2K/4K, and 12K/2K/4K.

4.3.2 Evaluation Protocol

F_1 value is calculated and evaluated for each positional relationship.

4.3.3 Results

Table 3 shows the experimental results (as the mean of five runs). The proposed PAS `twitter-pzero` model achieved the best performance in ZAR. The model pretrained with PZERO outperformed the model pretrained with Cloze. This suggests that prior learning by PZERO is linked to the acquisition of adaptive knowledge, which is consistent with the results of existing studies (Konno et al., 2021). The model pretrained with Twitter data performed better than did the model pretrained with Wikipedia data, especially showing large improvements with *exophora* (+2.2% on Wikipedia data, +1.7% on Twitter data).

4.4 Experiment 2: DC Model

We evaluated the complementation performance of the DC models within patterns (e) to (h) in Table 2. The DC model uses the results of the PAS model to output a sentence that completes for omissions appearing in the input sentence.

4.4.1 Dataset

We used JPersonaChat and JEmpatheticDialogues (Sugiyama et al., 2021) to evaluate the DC model. These datasets will also be used in Section 4.5.

JPersonaChat. JPersonaChat is a Japanese version of PersonaChat (Zhang et al., 2018) that assigns personas to two speakers and collects chat dialogues in which they learn more about each other. We split this dataset so that the numbers of dialogue pairs in the training/development/test sets were 50K/3K/4K. This corpus consists of persona description and dialogue pairs, but please note that we do not use persona descriptions in this work.

JEmpatheticDialogues. JEmpatheticDialogues is the Japanese version of EmpatheticDialogues (Rashkin et al., 2019), a dataset of utterances and corresponding empathic responses in emotional situations. We split this dataset so that the numbers of dialogue pairs in the training/development/test sets were 50K/3K/7K.

4.4.2 Evaluation Protocol

We performed human evaluations of the DC model performance, using 250 randomly sampled dialogues from the JPersonaChat and JEmpatheticDialogues test sets for each of the four models. Five evaluators were presented with two dialogue histories, one before and one after completion, and

ID	Model	ZAR				*dep*	All
		All	*intra*	*inter*	*exophora*		
(e)	PAS wiki-cloze	62.27	68.39	**44.63**	67.77	**94.17**	83.67
(f)	PAS twitter-cloze	62.21	68.04	40.68	70.34	94.15	83.73
(g)	PAS wiki-pzero	62.68	68.35	43.02	69.99	93.96	83.75
(h)	**PAS twitter-pzero**	**63.25**	**68.68**	42.07	**72.04**	93.81	**83.87**

Table 3: Automatic evaluation results by the PAS model (F_1).

ID	Model	Appropriateness
(e)	DC wiki-cloze	74.80% (187 / 250)
(f)	DC twitter-cloze	77.20% (193 / 250)
(g)	DC wiki-pzero	72.40% (181 / 250)
(h)	**DC twitter-pzero**	**84.80%** (**212** / 250)

Table 4: Human evaluation results of the DC model.

asked to judge whether the completion phrase and its position were appropriate. Each evaluator evaluated 1,000 data divided into five parts, 50 per model, for a total of 200 data for the four models. To ensure fairness, the dialogue histories completed by each model were shuffled before presentation to the evaluator, thus obfuscating which model completed which.

4.4.3 Results

Table 4 shows the experimental results. The proposed DC twitter-pzero model achieved the best performance in dialogue completion. For the model pretrained with the Cloze task, using Twitter data instead of Wikipedia data for pretraining improved the performance by 2.4% (from 74.80 to 77.20). In the model pretrained with the PZERO task, using Twitter data instead of Wikipedia data for pretraining improved the performance by 12.4% (from 72.40 to 84.80). This suggests that using Twitter data for pretraining the DC model contributes to improving the performance of dialogue completion. Furthermore, in Table 3, pattern (h) shows the best performance, suggesting a relation between the performance of dialogue completion and that of predicate-argument structure analysis.

4.4.4 Analysis

Table 5 shows cases of successful and unsuccessful dialogue completion for analysis. Examples 1 and 2 are successful completion cases. In Example 1, the NOM and ACC cases corresponding to "cause" are completed correctly, and in Example 2, the NOM and ACC cases corresponding to "help" are also completed correctly. Examples 3 and 4 are cases of failed completions. In the sentence in Example 3,

Example 1:	I ate oysters at a barbecue and {$\phi_{\mathrm{NOM}} \rightarrow$ ✓ **oysters** } caused {$\phi_{\mathrm{ACC}} \rightarrow$ ✓ **me** } to suffer from stomach pains and diarrhea all night long.
Example 2:	The other day a classmate was bullied and {$\phi_{\mathrm{NOM}} \rightarrow$ ✓ **I** } helped {$\phi_{\mathrm{ACC}} \rightarrow$ ✓ **him** } out.
Example 3:	I spent a little too much {$\phi_{\mathrm{ACC}} \rightarrow$ ✗ **on my credit card** } last month ... credit card.
Example 4:	I was having a lot of morning sickness and {$\phi_{\mathrm{NOM}} \rightarrow$ ✗ **morning shickness** } was lying on a bench in the supermarket and someone talked to me.

Table 5: **Examples of DC model completion results (translated from Japanese).** Highlighted text represents complemented words. ✓ indicates a correct completion, while ✗ indicates an incorrect completion.

the argument corresponding to "spend" should not be completed because inverted sentences occur in the utterance. Since this method judges whether to perform completion by looking at the front of the target predicate, the method could not complete sentences with inverted predicates. To perform a correct completion, it is necessary to devise a way to rewrite "I spent a little too much on my credit card last month" before inputting it to eliminate the inversion occurring in "I spent a little too much last month ... credit card." In Example 4, "I" is the correct answer, but "morning sickness" is incorrectly completed. This problem could only be solved by using as a clue the knowledge that morning sickness is a phenomenon, and appropriate dialogue completion was not possible for a problem that required such common-sense knowledge.

4.5 Experiment 3: RG Model

We evaluated the performance of patterns (a) to (h) in Table 2 in generating dialogue responses. The RG model uses BERT2BERT (Rothe et al., 2020), which uses BERT as both the encoder and decoder. Patterns (a) to (d) are the baseline models, and pat-

terns (e) to (h) are the proposed models applying the DCZAR framework. The baseline model uses the dialogue history (text before completion) contained in the dataset. The proposed model uses as input the dialogue history complemented by the DC model.

4.5.1 Dataset

The RG model is trained using dialogue history–response pairs, with only the dialogue history used as input for response generation during inference. We used JPersonaChat and JEmpatheticDialogues to fine-tune the RG model. `[SPK1]` and `[SPK2]` are added as special tokens. These special tokens are added immediately before the utterances of the two speakers in the dialogue history to make it easier for the model to distinguish between each speaker.

4.5.2 Evaluation Protocol

Automatic Evaluation. We used standard natural language generation metrics such as BLEU (Papineni et al., 2002), ROUGE-L (Lin and Och, 2004), DIST-N (Li et al., 2016), and BERTScore (Zhang et al., 2020a).

Human Evaluation. All evaluators evaluated all the 100 randomly sampled cases from the JPersonaChat and JEmpatheticDialogues evaluation sets for each of the four pretraining models, for a total of 400 cases. Three evaluators were presented with the dialogue history and two responses generated by two models (proposed method, baseline), and were asked to choose one or select *not sure* for evaluation criteria in a pair-wise comparison. The responses were evaluated in three dimensions: which was more *grammatical*, which was more *coherent*, and which was more *engaging*. To ensure fairness, the responses generated by each model were shuffled before presentation to the evaluators, making it impossible to distinguish which model generated which response. The final evaluation value was determined by a majority vote of the three evaluators.

4.5.3 Results

Automatic Evaluation. Table 6 shows the results of a single run of the automatic evaluation. We performed a permutation test for each proposed method and each baseline method. For BLEU-1, 3, 4 and ROUGE-L, the proposed method outperformed the baseline method, but there was no significant difference. Although the proposed method was expected to produce more coherent and engaging responses by compensating for predicate arguments, these automatic metrics were not necessarily appropriate, because their contribution was not expected to change the results of the word statistics.

Human Evaluation. Table 7 shows the results of human evaluation. * and ** indicate a significant difference with $p < 0.05$ and 0.01, respectively, by the chi-square test. Note that although this table shows the values after the majority vote, the values before the majority vote were used for the chi-square test. First, no models differed significantly in terms of grammaticality, but the RG `twitter-cloze`+DCZAR and RG `twitter-pzero`+DCZAR models exceeded the baseline. One possible reason for the lack of significant differences is that the number of N/A cases was higher than it was for the other perspectives. In terms of coherence, all models in which the DCZAR framework was applied showed significant improvements over the baseline model. In particular, the proposed RG `twitter-pzero`+DCZAR model shows a significant improvement as compared with the RG `twitter-pzero` model (from 38 to 62). This indicates that the use of dialogue history with explicit completion of omissions in the input contributes to coherence evaluations when generating responses. In terms of engagement, all models except the RG `wiki-pzero`+DCZAR model showed significant improvements over the baseline model.

4.5.4 Analysis

We analyzed the generated sentences in Table 8.

Why was there no significant difference in grammaticality scores between the baseline and the proposed method? This was possibly due to the higher number of N/A results as compared with the other perspectives. Dialogue 1 is an example where three evaluators selected *not sure* and the response was classified as N/A. In this example, although the two models generated responses with different content, neither response was grammatically incorrect, and the decision may have been difficult in this case.

Does the proposed method contribute to improved coherence? Dialogue 2 is an example evaluated as contributing to the generation of a more coherent response by the proposed method. In the dialogue history of Dialogue 2, there are

ID	Model	BLEU				ROUGE-L	DIST		BERT
		1	2	3	4		1	2	Score
(a)	RG$_{\texttt{wiki-cloze}}$	25.29	6.14	2.02	0.69	9.57	**12.20**	**29.32**	69.70
(e)	+ **DCZAR (ours)**	24.50	5.65	1.78	0.55	**14.50**	11.72	28.50	69.45
(b)	RG$_{\texttt{twitter-cloze}}$	25.65	**6.50**	2.10	0.70	9.79	12.04	29.02	69.84
(f)	+ **DCZAR (ours)**	**25.72**	6.16	1.96	0.66	11.65	12.14	28.95	69.73
(c)	RG$_{\texttt{wiki-pzero}}$	25.59	6.31	2.09	0.72	13.72	12.09	28.96	**69.90**
(g)	+ **DCZAR (ours)**	25.45	6.08	1.96	0.63	6.49	12.06	29.14	69.75
(d)	RG$_{\texttt{twitter-pzero}}$	25.00	6.08	2.02	0.69	11.99	12.17	29.31	69.74
(h)	+ **DCZAR (ours)**	25.50	6.17	**2.11**	**0.73**	9.41	11.72	28.54	69.77

Table 6: Automatic evaluation results of the RG model.

ID	Model	gramm-atical	coherent	enga-ging
(a)	RG$_{\texttt{wiki-cloze}}$	**30**	45	44
(e)	+ **DCZAR (ours)**	28	**54****	**55****
	N/A	42	1	1
(b)	RG$_{\texttt{twitter-cloze}}$	30	43	46
(f)	+ **DCZAR (ours)**	**34**	**57****	**54****
	N/A	36	0	0
(c)	RG$_{\texttt{wiki-pzero}}$	**34**	45	**51**
(g)	+ **DCZAR (ours)**	33	**52****	47
	N/A	33	3	2
(d)	RG$_{\texttt{twitter-pzero}}$	32	38	41
(h)	+ **DCZAR (ours)**	**38**	**62****	**59****
	N/A	30	0	0

Table 7: **Human evaluation results of the RG model.** N/A shows the number of cases in which *not sure* was selected by a majority vote and in which the evaluation was split among the three evaluators.

Dialogue 1	
Speaker A:	I'm worried about my parents' retirement.
Speaker B:	{ϕ_{NOM} → **Our parents' retirement** } is a concern.
Speaker A:	I live far away from my parents and I don't know what they think.
RG$_{\texttt{wiki-cloze}}$:	We are concerned.
+ **DCZAR (ours)**:	It makes us nervous. We should think of something.
Dialogue 2	
Speaker A:	Being a housewife, there're so many times I thank my husband.
Speaker B:	I'm sure he's a kind spouse.
Speaker A:	Regarding the year-end gift, {ϕ_{NOM} → **my husband** } gives {ϕ_{DAT} → **gifts** } to {ϕ_{ACC} → **relatives** } on his wife's side as well as relatives on his side without any complaint.
RG$_{\texttt{twitter-pzero}}$:	**Your wife** is kind.
+ **DCZAR (ours)**:	**Your husband** is a great guy.
Dialogue 3	
Speaker A:	I told my mother that I wasn't feeling well and {ϕ_{NOM} → **she** } drove me to the hospital.
RG$_{\texttt{twitter-cloze}}$:	That would be great.
+ **DCZAR (ours)**:	Your mother is very kind.

Table 8: **Generation examples (translated from Japanese).** Highlighted text shows words filled in by our proposed model.

many nouns that express the role of a person, such as husband, spouse, wife, and housewife, and the arguments corresponding to "give" is omitted, making it difficult to guess who is being talked about in the sentence. The RG$_{\texttt{twitter-pzero}}$ model misinterprets the dialogue as being about the wife, and so generates an inappropriate response. By contrast, the proposed RG$_{\texttt{twitter-pzero}}$+DCZAR model correctly interprets the dialogue as being about the husband, so it generates an appropriate response. This suggests that the proposed method explicitly supplements the arguments corresponding to the predicate, which improves coherence.

What are the characteristics of engaging responses? The proposed method evaluated Dialogue 3 as contributing to the generation of a more engaging response. This example shows that the responses evaluated as more engaging are those that are more specific and coherent. Therefore, we hypothesized that the more coherent the response, the more engaging the response, and analyzed correlations among the indicators. The correlation coefficient for grammatical correctness and engagement was 0.223 and for coherence and engagement was 0.850. This indicates a strong correlation between coherence and attractive responses, and that people tend to want to continue dialogue with those who are consistent in their communication. Fleiss' Kappa (Fleiss, 1971), a measure of agreement among the evaluators, was calculated to be 0.095 for grammatical correctness, 0.287 for coherence, and 0.214 for engagement. The human evaluations indicated that the DCZAR framework contributed to the generation of more coherent and engaging responses.

5 Conclusion

We proposed the DCZAR framework, which explicitly completes omitted information in the dialogue history and generates responses from the completed dialogue history. Experimental results showed that the DCZAR framework can generate more coherent and engaging responses.

Limitations

We outline some potential limitations of our work below. First, extending to other languages requires pretrained models and datasets for each task (PAS, DC, RG) in that language. Also, our results do not necessarily guarantee the same results in languages other than Japanese. As we mentioned in Section 4.4.4, dialogue completion does not work well with inverted sentences and sentences that require common-sense knowledge as completion cues. Extending the DC model to handle such cases is a task for future work.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP22H00804 and JP21K18115; JST AIP Acceleration Research JPMJCR22U4, Japan.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 2206–2240.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

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

Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu, and Dilek Hakkani-Tur. 2021. Multi-sentence knowledge selection in open-domain dialogue. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 76–86, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuji Matsumoto. 2010. Annotating predicate-argument relations and anaphoric relations: Findings from the building of the naist text corpus. *Journal of Natural Language Processing*, 17(2):25–50.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Ryuto Konno, Shun Kiyono, Yuichiroh Matsubayashi, Hiroki Ouchi, and Kentaro Inui. 2021. Pseudo zero pronoun resolution improves zero anaphora resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3790–3806, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 758–766, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Robert C. Stalnaker. 1978. Assertion. *Pragmatics*, pages 315–332.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. arXiv:2109.05217.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th Conference on Neural Information Processing Systems*, pages 3104–3112.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. 2008. A Japanese predicate argument structure analysis using decision lists. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 523–532, Honolulu, Hawaii. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language models for dialog applications. arXiv:2201.08239.

Masato Umakoshi, Yugo Murawaki, and Sadao Kurohashi. 2021. Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1920–1934, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 93–107, Dublin, Ireland. Association for Computational Linguistics.

Souta Yamashiro, Hitoshi Nishikawa, and Takenobu Tokunaga. 2018. Neural Japanese zero anaphora resolution using smoothed large-scale case frames with word embedding. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

	Articles	Sentences	Predicates
Train	1,751	24,283	68,753
Dev	480	4,833	13,882
Test	696	9,284	26,379

Table 9: Statistics for NAIST Text Corpus 1.5.

		dep	*intra*	*inter*	*exophora*
Train	**NOM**	37,678	11,556	7,518	11,516
	ACC	24,997	1,803	928	128
	DAT	5,855	360	278	60
Dev	**NOM**	7,550	2,556	1,766	1,917
	ACC	5,107	394	166	32
	DAT	1,637	112	99	28
Test	**NOM**	14,254	4,770	3,342	3,721
	ACC	9,532	786	358	55
	DAT	2,547	211	140	54

Table 10: Distribution of arguments in NAIST Text Corpus 1.5.

A Ethical Considerations

Since the dialogue response generation model uses large-scale data from websites (e.g., Wikipedia, Twitter) during pretraining, it may generate responses that contain implicit biases and offensive content. We will incorporate mechanisms to reduce harmful responses and build a safe and ethically robust dialogue system in the future.

B Details of Scientific Artifacts

B.1 Dataset

Wikipedia. We used a publicly available data dump of Japanese Wikipedia jawiki-latest-pages-articles.xml.bz2.

Twitter. We used preprocessed tweets collected through the Twitter API[1] to pretrain the model. We used all tweets by 3,702 users with tweet histories ranging from 10K to 50K postings, sorted in chronological order.

NAIST Text Corpus 1.5. We used NAIST Text Corpus (NTC) 1.5 to test the performance of the PAS model. NTC is a corpus of newspaper articles and editorials with information such as relations between predicates and surface cases. Table 9 shows the NTC statistics, and Table 10 shows the distribution of NTC arguments.

[1] https://developer.twitter.com/en/products/twitter-api

B.2 Model

In this work, we used HuggingFace Transformers[2] version 4.21.0 (Wolf et al., 2020), and weights of bert-based-japanese-whole-word-masking[3] provided in transformers were used as initial parameters for the pretrained model.

B.3 Metric

For the BLEU[4], ROUGE-L[5], and BERTScore[6] implementations, we used publicly available code from Huggingface.

B.4 Software

We used MeCab 0.996[7] , a Japanese morphological analyzer, and CaboCha 0.69[8], a Japanese dependency analyzer, to preprocess the dataset. We will release our code publicly available.

B.5 License

As for the datasets, Japanese Wikipedia is made available under the CC BY-SA 3.0 license, NAIST Text Corpus 1.5 is released under a Revised BSD License, and JPersonaChat and JEmpatheticDialogues are licensed for the purpose of evaluating the model performance, but not for providing dialogue services themselves. MeCab is available under three licenses (BSD, LGPL, and GPL), and CaboCha is released under the Revised BSD License. The bert-based-japanese-whole-word-masking model is available under the CC BY-SA 3.0 license. Since both licenses allow use for research purposes, the use of these artifacts is valid for this work.

C Details of Experiments

C.1 Software and Hardware

We used Python 3.8, PyTorch 1.12.1, and HuggingFace Transformers 4.21.0. All experiments were performed using two NVIDIA A100 80 GB GPUs for model pretraining and one NVIDIA A100 80 GB GPU for fine-tuning. The pretraining time was about six days per model, fine-tuning for

[2] https://github.com/huggingface/transformers/

[3] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[4] https://github.com/huggingface/datasets/blob/main/metrics/bleu/bleu.py

[5] https://github.com/huggingface/datasets/blob/main/metrics/rouge/rouge.py

[6] https://huggingface.co/spaces/evaluate-metric/bertscore

[7] https://taku910.github.io/mecab/

[8] https://taku910.github.io/cabocha/

Pretraining	
Mini-batch Size	2048
Max Learning Rate	1.0×10^{-4} (Cloze)
	2.0×10^{-5} (PZERO)
Learning Rate Schedule	Inverse square root decay
Warmup Steps	5,000
Number of Updates	30,000
Loss Function	Cross entropy (Cloze),
	KL divergence (PZERO)
Fine-tuning of the PAS model	
Mini-batch Size	256
Max Learning Rate	5.0×10^{-5}
Number of Epochs	20
Loss Function	KL divergence,
	Cross entropy (*exophora*)
Fine-tuning of the RG model	
Mini-batch Size	1,024
Max Learning Rate	5.0×10^{-5}
Number of Epochs	30
Max Sequence Length	512 (encoder)
	128 (decoder)

Table 11: List of hyperparameters.

the PAS model was about nine hours per model, and fine-tuning for the RG model was about seven hours per model.

C.2 Hyperparameters

Table 11 lists the hyperparameters used in this study.

D Details of Human Evaluation

Since the human evaluations in these studies (Section 4.4.2 and Section 4.5.2) did not require expert knowledge of linguistics, we recruited eight Japanese undergraduate and graduate student evaluators from within our laboratory. We informed the evaluators of the purpose of this evaluation in advance and obtained their consent. Table 12 shows an English translation of the instructions given to the evaluators, who were paid for their time in accordance with university regulations.

E Experimental results in Japanese

The experimental results of the Japanese versions of Table 5 in Section 4.4 and Table 8 in Section 4.5 are shown in Table 13 and Table 14, respectively.

Human evaluation of the DC model

Task Info:
I am researching a dialogue response generation system, and as part of that research,
I need to evaluate the system's completion performance.
You are to read a sentence upon which a completion operation has been performed
and evaluate the appropriateness of both the wording and the position,
using a binary value of 1 (appropriate) or 0 (not appropriate).
Note that there may be typos in these sentences.
One example is the incorrect *Kodomo ga chiisai-no-shi* (which should be *Kodomo ga chiisai-shi*).
Please do not consider typos that are present in the source texts in your evaluation.

Example:
(a) When both the completion phrase and the completion position are appropriate.
`Speaker A: I took my summer suit to the cleaners.`
`Speaker B: Well done! When will` **`the suit`** `be ready?`
– The correct word and the position of the complement are both appropriate, so select 1.

(b) When the complementing phrase is not appropriate, but the completion position is appropriate.
`Speaker A: Even if I had a boyfriend, I would break up with him right away.`
`Speaker B: Maybe` **`I`** `just haven't met the man of my dreams yet.`
– The completion position is appropriate, but the appropriate complement phrase is `You` instead of `I`. In such a case, select 0.

Human evaluation of the RG model

Task Info:
I am researching a dialogue response generation system, and as part of that research,
I need to evaluate its response performance.
You will be given multiple dialogue contexts and two responses.
Please compare Responses 1 and 2 and evaluate which is more grammatical,
which is more coherent, and which is more engaging.
grammatical: Which response is more grammatical and fluent in Japanese
(ignoring the dialogue history)?
coherent: Which response is more coherent, considering the dialogue history?
engaging: Which response is more engaging, and which response makes you want to continue
the dialogue with the person to whom you are talking?
Select 1 if Response 1 is better, select 2 if Response 2 is better, or select 3 if you are unsure
which response is better. Use choice 3 as sparingly as possible.

Example:
You will evaluate the following Responses 1 and 2 from three perspectives.
`Speaker A: I bought a gaming console.`
`Speaker B: Did you buy your first game?`
`Speaker A: Yes, it's surprisingly interesting.`
—
`Response 1: What game are you playing?`
`Response 2: I see. Good for you.`

Table 12: Instructions for evaluators (translated from Japanese).

Example 1:	バーベキューで牡蠣を食べたら、 牡蠣が 見事に当たって 一晩中 私が 腹痛と下痢に苦しみました。
Example 2:	この前、クラスメートがいじめられてたから、 私が クラスメートを 助けたんだ。
Example 3:	先月 カードを ちょっと使いすぎちゃったんだよなクレジットカード。
Example 4:	つわりが酷くて つわりが スーパーのベンチに横たわってたら いろんな人が大丈夫ですか？って声かけてくれたよ。

Table 13: **Examples of DC model completion results in Japanese.** Highlighted text represents complemented words.

	Dialogue 1
Speaker A:	親の老後が不安。
Speaker B:	この世代になると、 老後が 不安になってくるよね。
Speaker A:	そうなんだよね。離れて暮らしてるし、親の意向も分からないしでね。
RG$_{\text{wiki-cloze}}$:	心配だよね。
+ DCZAR (ours):	不安になるよね、何か考えておかないとね。
	Dialogue 2
Speaker A:	専業主婦をしてると、夫に感謝する場面がとても多いわ。
Speaker B:	きっと優しいご主人なんだろうね。
Speaker A:	お歳暮も、夫の親戚と同じように、妻側の親戚にも 嫌な顔ひとつせずに 夫が 歳暮を 親戚に 贈ってくれるのよ。
RG$_{\text{twitter-pzero}}$:	貴方の奥さん優しいなあ。
+ DCZAR (ours):	それは素晴らしい旦那さんね。
	Dialogue 3
Speaker A:	母に具合が悪いことを伝えたら、 母が 病院まで車で送ってくれました。
RG$_{\text{twitter-cloze}}$:	それはありがたいね。
+ DCZAR (ours):	お母さま、優しいですね。

Table 14: **Generation examples in Japanese.** Highlighted text shows words filled in by our proposed model.