Evaluation of Active Generation of Interlocutor Profiling Sentences from Utterances and Their Implicit Context

Shinji Muraji[1], Rafal Rzepka[1] and Toshihiko Itoh[1]

[1]Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan

Abstract

Recently, it has been reported that the quality of chats can be improved by utilizing information about what kind of persona the chat interlocutor and what kind of personality the system itself mimics in order to generate responses using LLMs. In these studies, personality data is stored in a form of statements describing interlocutor's profile. In order to control LLM using profile describing sentences in actual chats, it is not enough to prepare them in advance, but it is necessary to actively add them through the chats. However, not enough research has been conducted on generating the profiling sentences from the utterances in the chat dialogs. In particular, there are no studies on the generation of profiling sentences that can only be generated with context the context. Therefore, in this study, we propose a task to generate profiling sentences from the target utterances based on context in a chat dialog in Japanese, and create a dataset for this task. Experiments on the created dataset and analysis show that LLM can generate profiling sentences while taking the context into account.

Keywords

Chat system, Profiling sentence generation, Persona modelling

1. Introduction

In recent years, the performance of large language models (LLMs) has dramatically improved, leading to significant advancements in dialog systems. However, in case of chat, dialog systems based solely on language models suffer from limitations due to the inability to remember users in the long term and inconsistent utterances in simulating personality. Related research [1, 2] has shown that the quality of dialog in a chatting system can be improved by maintaining personality information, such as what kind of person one is talking to and what kind of personality one imitates, separately from the dialog history, and utilizing this information for response generation . In studies on long-term memory, maintaining personality information separately from the dialog history is useful for remembering the personality of the conversation user enabling appropriate utterances tailored to that interlocutor [2]. In the study of utterance consistency, it is useful to learn the relationship between a personality and an utterance and to map a candidate utterance to a coherent utterance [3]. Many studies [1, 4] have employed personality information in the form of natural language list (hereafter referred to as profiling sentences) and have shown its effectiveness.

In order to utilize profiling sentences for response generation in actual conversations, it is essential to prepare profiling sentences that express the personality of the interlocutor (hereinafter referred to as *user profiling sentences*) and profiling sentences that represent the personality that the system should assume as its own personality (hereinafter referred to as *system profiling sentences*), which have been collected during previous dialogs. Previous studies have either prepared profiling sentences in advance [1] or substituted sentences extracted by rules [5] for profiling sentences, Although user profiling from utterances is important for both dialog consistency and user adaptation, research on this topic is rather scarce.

In previous research on improving the consistency of the system's own dialog by using system profiling sentences, about five profiling sentences are prepared as a part of the personalities commonly used in chats before starting a chat and are added to the input for dialog generation [1]. However, methods that use only the pre-defined profiling sentences as the system profile do not control personalities

LaCATODA 2024: The 9th Linguistic and Cognitive Approaches to Dialog Agents Workshop, Nov 19, 2024, Kyoto, Japan

shinjimuraji@ist.hokudai.ac.jp (S. Muraji); rzepka@ist.hokudai.ac.jp (R. Rzepka); t-itoh@ist.hokudai.ac.jp (T. Itoh)

Table 1
Example of human profiling sentences inferred from utterances (translated from Japanese)

A’s utterance	I will be 20 years old on the big day. I’m a Capricorn.
B’s utterance	You are young. I work in the apparel industry. Are you a college student?
A’s utterance (target utterance)	Yes. I’m graduating soon, so I was wondering what I should do in the future. I like fortune-telling and stuff like that. I thought it would be difficult to make it a profession. What kind of things do you do in the apparel industry?
profiling sentence	A is a college student. A is about to graduate from college. A likes fortune telling. A is worried about finding a job.

beyond these sentences, and thus cannot maintain consistency over a long period of time. It is not realistic to prepare in advance information about all aspects of a person’s personality as system profiling sentences to maintain consistency, hence the system needs to make sure that the utterance it is about to generate is consistent with past utterances.

In addition, it is desirable for the system to be able to recognize user profiling sentences from utterances during chats, since it would be a great burden to have the users themselves prepare their own profiling sentences in advance.

Thus, both long-term memory and utterance coherence research require the recognition of individuality from utterances. We address this issue by generating profiling sentences from utterances as a task of personality recognition.

We propose a task to generate profiling sentences with LLMs that take into account the context of chats, and construct a dataset by LLM generation and human consistency checks. Table 1 shows examples of sentences inferred by human from utterances and their implicit context. In addition, we propose a method to train LLM and generate profiling sentences on that dataset, and report on the automatic and manual evaluation of the generated results. Furthermore, we confirm the importance of considering the context of the chat by manually checking the dataset itself and the output of the trained profiling sentence generation model are context-sensitive.

The contributions of this paper are as follows:

- Proposed a task to generate profiling sentences considering the implicit chat context
- Proposed method for training profiling sentences generation with LLM
- Identified differences between automatic and manual evaluation in profiling sentences generation
- Confirmed that profiling sentences generation by LLM is difficult to take into account the chat context

2. Related Work

Currently, there are three main methods for linking utterances to profiling sentences, but the existing methods are not sufficient for the task of recognizing profiling sentences in chit-chat utterances.

First, there is an approach that uses classification to recognize whether an utterance corresponds to a profiling sentence. For example, some studies [6] treat the task of determining whether an utterance contradicts a profiling sentence as a classification problem for implication relations. Other studies [1, 7] focus on preparing several profile sets, each consisting of about five sentences, and then selecting the profile set that corresponds to the utterance. However, these methods are difficult to use because the task is to classify the relationship between the profiling sentences prepared in advance and the utterances, and to perform the classification in actual chats where any topic may appear, all kinds of profiling sentences must be prepared in advance.

Second, there is an approach to extract profiling sentences from dialog by rules. For instance, [5] searched Reddit comments for sentences that fit several rules, such as “contains either the word I or my,” and used them as a substitute for the commenter’s profiling sentence. Although such a method for extracting literal profiling sentences is effective to a certain extent, it may extract sentences that require contextual processing when applied to chat dialogs. (For example, “A is a college student.” in Table 1 cannot be extracted from target utterance only.) In addition, the creation of rules is labor-intensive, and the same rules may not be applicable to some languages. In particular, the subject *I* is often omitted in Japanese uterances, which is the target of this study, and many profiling sentences would be overlooked if this rule is applied.

Third, there is a method to generate profiling sentences from dialogs using LLM. However, existing research [4] ignores the context of the previous chats and does not perform human evaluation of the generated profiling sentences.

3. Profile generation task

During a chat, humans recognize and update information about the personality of the person they are talking to each time they receive the other person’s utterance. Regarding the recognition part, humans can to some extent write down the information of the other party as short sentences for each utterance. It is important to note that the profiling sentences written out by the human are not extracted solely from the target utterance, as shown in Table 1. Our goal is to enable the system to extract and record the interlocutor’s information as profiling sentences from the utterance, similarly to human recognition. As in previous studies, we do not define the exact nature of a profiling sentence. Instead, we treat all *sentences that describe personality or personal facts* as profiling sentences, consistent with their usage and validation in prior research. It would be preferable generate human-like profiles in the system, but the annotation becomes very complex when trying to rigorously create such tagged dialog data. Therefore, this paper focuses on demonstrating the importance of generating profiling sentences by LLM using context. In this section, we summarize the differences between the ideal task we originally wanted to perform, the conventional task, and the proposed task.

3.1. Problems of previous studies

First, we describe the ideal task. Let Δp_t be one of the profile sentences of the interlocutor’s profiling sentence can be inferred from the tth utterance u_t of the dialogue, and let $\Delta P_t = \{\Delta p_t^1, \Delta p_t^2, \Delta p_t^3 ...\}$ be all the profile sentences from utterance u_t. In addition, $P_t = \{p_t^1, p_t^2, p_t^3 ...\}$ be all interlocutor’s profiling sentences collected from the previous dialogs, updated by ΔP_t after receiving u_t.

Note that some profiling sentences are updated, including the place of residence or hobbies, and that P_n is not the union set of ΔP_t from $t = 1$ to $t = n$.

Let $C_t = \{c_1, c_2, c_3 ...\}$ be the context other than the target utterance necessary to generate the information about another person known from the utterance, then the profiling sentence generated by model M_{human}, which has (theoretically) human-like capabilities, can be $\Delta P_t = M_{human}(u_t, C_t)$. The ideal profiling sentence generation is to obtain M_{human} and generate human-like profiling sentences. However, C_t can assume a variety of elements, which is problematic when creating a dataset. In the case of a human predicting an interlocutor’s profiling sentence from utterance, the C_t can be assumed to be the profile of a dialog interlocutor already known, the dialog history, or any other concept that is shared with the interlocutor. The elements of C_t in this paper are the interlocutor’s profiling sentence P_{t-1} held before receiving an utterance and the history of utterances $D_t = \{u_1, u_2, u_3 ..., u_{t-1}\}$. When starting to talk, C_1, D_1, and P_0 are empty sets. In this study, we do not update the profile, but for the sake of explanation, we denote the profile update model as MP. The formula for profile update is $P_t = MP(\Delta P_t, P_{t-1})$. In summary, ideally the profiling sentence should be extracted and updated as follows:

- Extraction: $\Delta P_t = M(u_t, C_t) = M(u_t, P_{t-1}, D_t)$
- Update: $P_t = MP(\Delta P_t, P_{t-1})$

Table 2
Examples of utterances with no recognizable profiling sentences from a target utterance and a conversation initiation utterance (translated from Japanese)

Utterances	profiling sentence(s)
A's utterance: I guess it's a compliment, but do some people say it in a sarcastic way? I am insensitive, so I don't notice such things. B's utterance: I see. You seem like a very open-minded and kind person. I hope you will talk to me again! A's utterance (target utterance): Yes, let's talk again. Thank you very much!	(empty)
A's utterance (target utterance): Hello. I am a graduate student. What are you doing now?	A is attending graduate school.

It is most desirable to create a dataset by storing P_t and ΔP_t for each u_t, but in reality, problems such as the burden and cost of annotators appear. The longer the conversation is, the more profiling sentences P_t are collected for each chat, and it is not easy to annotate each utterance while keeping track of all of profiling sentences.

On the other hand, the task setting of the profiling sentence generation in the previous study [4] can be expressed as $\Delta P_t = M(u_t)$, which is the ideal task minus the context information C_t. However, there is also persona information for which contextual information is essential. An example that requires contextual information is shown in Table 1. In a previous study, [8] analyzed how much contextual information is needed, and they found that humans use context in about 15% of profiling sentences when inferring profiling sentences. Therefore, in a task setting that excludes contextual information, even humans have an upper limit of 85% recall, which means that one out of every six profiling sentences is impossible to be discovered in the first place.

3.2. Task to generate profiling sentences from context and utterance

We propose a task to generate profiling sentences from a target utterance while considering the two previous utterances as context. Here, the P_t profiling sentence is an important element, but we ignore it in this study. In our task setting, it is crucial to effectively utilize the two preceding utterances as context C_t for generation. In other words, with the utterance history $D_t = \{u_{t-1}, u_{t-2}\}$, our proposed task can be expressed as follows.

$\Delta P_t = M(u_t, C_t) = M(u_t, D_t) = M(u_t, u_{t-1}, u_{t-2})$

This task aims to generate zero or more known profiling sentences ΔP_t from the target utterance u_t using contextual information u_{t-1},u_{t-2}. Table 1 shows examples of utterances with 4 profiling sentences, Table 2 shows examples of utterances with 0 and 1 profiling sentences. In the previous study, utterances that humans inferred to be without profiling sentences were not included in the target utterances. However, in this study, we included them to simulate a more realistic chatting scenario. Since humans can also infer implicit contextual information u_{t-1} and u_{t-2} to predict profiling sentences from target utterances u_t, the results are compared with the results of a human performing $\Delta P_{t\_gold} = HUMAN(u_t, u_{t-1}, u_{t-2})$ in the same task.

We create a dataset for this task and provide a benchmark. We also compare and analyze ΔP_t and ΔP_{t_gold} in the proposed task to confirm the importance of context in profiling sentence generation.

4. Data collection

In this section, we first describe the chat data we use to build our dataset, and then explain how we create profiling sentences from the chat data and ensure their quality.

4.1. Original chat data

We extend an existing chat dataset by creating a dataset that links target utterances with context-sensitive profiling sentences. The original dataset for the extension is the JPersonaChat dataset [9]. This dataset is a Japanese version of the PersonaChat dataset [1]. This dataset is a collection of data from a dialogue between crowdworkers who were given a five-sentence pre-profiling sentence and who play their asigned roles. In order to play their assigned role, crowd workers tend to make statements related to their role. Therefore, we decide to utilize it because we considered it suitable for linking utterances to profiles. It is important to note that our goal is to predict the profiling sentences that will be inferred from utterance, but the profiling sentences given in advance are not necessarily the profiling sentences that will be recognized from utterance. While playing the role, the crowdworker may add a profiling sentence to the utterance that has not been given in advance. There are also cases where the profiling sentences given in advance are not used in the target utterance. Therefore, we do not use the profiling sentences for the roles given to the crowd workers in the original dataset. We use only the dialogues as the chat dataset.

4.2. Interlocutor profiling sentence creation

As described in the task setup section, we add the two most recent utterances in the utterance history, including the interlocutor utterance, to a single target utterance as contextual information. We use a subset of the original dataset divided by dialogue. The target dialogs for extraction are 500 dialogs randomly selected from the JPersonaChat dataset, which contains 5,448 utterances. All utterances in the obtained subset dialogs are considered as target utterances, and each target utterance and the two utterances immediately preceding it are considered as one case of data. Here, we counted the number of utterances required by one of the authors to infer profiling sentences for 100 target utterances randomly selected from the original chat data, and found that there was only one target utterance that required three or more utterances of context, so we set the number of utterances used for context to two. If the target utterance is within two utterances from the start of the conversation, the entire dialogue history is added. We want to obtain corresponding profiling sentences for each target utterance. However, it takes a lot of effort to obtain a comprehensive and accurate profiling sentences all by hand. The method in which a single annotator infers a profiling sentence for a single target utterance is unreliable. The method in which multiple people check the profiling sentences inferred by one person is considered to be more accurate, but it lacks comprehensiveness because it misses profiling sentences that were not recognized by the annotator who made the inference. Ideally, the profiling sentences written by several people should be merged, and the merged profiling sentences should be checked by several people, but this is very costly. Therefeore, we use LLM to create profiling sentence dataset to decrease the costs. This method of generating data using a language model and manually checking it is often used in recent years [10], and although limitations exist in the capabilities of LLMs, it is an efficient way to create data. The profiling sentence is intended for use in LLM response control and profile updating, which are subsequent processes, thus maintaining its accuracy is crucial. Therefore, LLM is used to write out as many different profiling sentences as possible, and human verification is used to create the dataset with guaranteed accuracy.

4.3. Generate candidate profiling sentences using LLM

This section describes the process of writing profiling sentences from utterances using LLM. We use *gpt-4-0613* as the LLM when creating the dataset. As a preliminary experiment, we asked LLM to infer profiling sentences using several prompts, and the results with instructions only (zero-shot) were more comprehensive than the results generated with examples of correct answers (few-shot). Therefore, we use a zero-shot setup for LLM inference. the LLM is given a prompt that is a concatenation of the instruction, two prior utterances from the dialog history, and the target utterance, and is asked to generate as many profiling sentences as possible for all the target utterances. The profiling sentences include utterances for which no profiling sentences exist, but if they do not exist, they are output as *none*.

As noted in the task description section (2.2), the target utterances for this study include those in which humans do not infer any profiling information. However, adding target utterances without profiling sentences to the dataset creates a situation where the linkage between them is no longer 1-to-many, rendering some existing automatic evaluation methods unusable. For a more accurate evaluation, the dataset is checked manually for validity. The human evaluation also compares the results of human inferences $\Delta P_{t\_gold} = HUMAN(u_t, u_{t-1}, u_{t-2})$ on a small number of utterances (100).

Table 3
Example of an annotator's judgment(translated from Japanese)

Utterances	profiling sentence	judgment
A's utterance: That's right. The other day, I was approached by a man because I have a distinctive face, but he seemed to be put off by my gray hair. B's utterance: Oh, that's shocking! Why don't you dye your hair like me, with a little gray in it? A's utterance (target utterance): I don't think I could go that far. But there are probably quite a few people like that in the Kinki region where I live.	A lives in the Kinki region.	*correct*

4.4. Evaluation of profiling sentence candidates

Next, to ensure the accuracy of the profiling sentences generated by the LLM, the correspondence between the target utterance and the profiling sentences is manually checked. This annotation work is done by a crowdworker. In order to check whether each profiling sentence can be inferred from the target utterance, we first divide a "target utterance, two prior utternaces" (hereafter referred to as "utterance group") and a "profiling sentence" to achieve one-to-one alignment. Annotators are assigned to evaluate pairs of *utterance groups* and *profiling sentences.* The information in the profiling sentences, derive from the target utterance, is assessed by three raters using three categories: *correct, possibly correct*, and *incorrect.* An example of an annotator's decision is shown in Table 3. A majority vote is used as the final decision, and in the case of a split decision by all workers, an intermediate label, *possibly correct,* is adopted. Note that our goal is to generate profiling sentences derived from the target utterance, thus any profiling sentences that are mentioned only in the utterance history and are unrelated to the target utterance are judged as *incorrect.* profiling sentences with incorrect Japanese or profiling sentences that can be applied to any utterance (e.g., "I can speak the language") are also judged to be *incorrect.* The total number of profiling statements inferred by the LLM is 16,971, of which 9,475 (55.83%) are judged *correct,* 1,763 (10.39%) are judged *possibly correct,* and 5,733 (33.78%) are judged *incorrect.* Data annotated as *incorrect* are not used in this experiment. The inter-annotator agreement (three-way average of weighted kappa coefficients) is 0.557, indicating moderate agreement. This suggests that there are some individual differences in what is perceived as a profiling sentence. After the annotation is completed, the dataset is created by reverting to data for each target utterance for correct and potentially correct "utterance groups" paired with "profiling sentences". When processing the data for each target utterance, the profiling sentences are left blank for target utterances that don't have any profiling sentences associated with them, as shown in the Table 2. The dataset we created is publicly available [1] .

[1] https://github.com/shingetsu-ak/generation-of-interlocutor-profiling

5. Context-sensitive profiling sentence generation experiment

We train models and conduct experiments to validate the datasets we create. In this section, we describe the details of model construction and training, evaluation method, experimental results, and an analysis of contextual influences.

5.1. Experimental settings

For training, the dataset is randomly shuffled for each target utterance and divided into subsets of training, validation, evaluation = (8:1:1) by the number of target utterances. Although it is possible for a target utterance to appear in the utterance history of other target utterance, since our task is to map target utterances to profiling sentences, we have separated them in this way because we believe that training, validation, and evaluation should be done on unique target utterances.

Although the previous study [4] on profiling sentences, uses training and evaluation by concatenating profiling sentences, there are several possible problems with simple concatenation of profiling sentences. First, inference by the language model causally predicts tokens sequentially, but when predicting multiple profiling sentences, later predictions may be influenced by earlier ones. When multiple profiling sentences are obtained from a single utterance, it is unclear whether one profiling sentence should be generated using another profiling sentence as the addition to an utterance. Therefore, we propose an alternative method to training simple concatenations: training utterance groups and profiling sentences on a one-to-one basis. For our experiments, we created models for both of these profiling sentence generators and compared their performance.

5.1.1. Model training details

In both the method for training simple concatenations (concat) and the method for training utterances and profiling sentences on a one-to-one basis (profile-wise), profiling sentences are generated using a Transformer-based decoder, following the approach in the previous study [4], and causal language modeling (CLM) as the objective function. Specifically, the model is trained by LoRA fine-tuning of *cyberagent/open-calm-7b*, a Japanese open source LLM[2]. LoRA fine-tuning uses the target utterance group as input to output profile sentences. No instructions are used. As hyperparameters, the learning rate is set to $5e-5$, the rank r of LoRA to 32, the weight decay to 0.01, the number of warm-up epochs to 1, AdamW is used for optimizer, and the batch size is set to 8. The model with the lowest loss out of 10 epochs is used as the best model of each methodss. When testing the model, top-p sampling is employed, with $p = 0.95$. As Ribeiro et al., we do not compute the loss during training for utterance groups, but only for the generation of profiling sentences.

We propose an one-to-one training method that learns a profiling sentence from a target utterance, but with this method the model can only generate one profiling sentence per inference. Therefore, in order to obtain multiple profiling sentences in one utterance, it is necessary to let the model infer multiple times and remove the same profiling sentence from the generated results. Therefore, to ensure that the diversity of the learned profile sentences is reflected in the generation of the profile sentences, they are generated 10 times. Since many of the 10 generated profiling sentences are semantically similar, the semantic similarity is measured by sentenceBERT [11] , and those exceeding the threshold value (experimentally set to 0.8 in this case) are integrated by eliminating them as duplicates. If the model outputs nothing at least once out of 10 times, the absence of a profiling sentence is given priority and the other outputs are discarded.

5.1.2. Generative models compared

We compare models that learn profiling sentences concatenated together, and models that learn utterance groups and profiling sentences one-to-one. profiling sentences that are confusing to humans may

[2] https://huggingface.co/cyberagent/open-calm-7b

confuse the model during training, thus we created a model that uses profiling sentences that the crowdworker judged to be “possibly correct” during dataset construction and a model that does not use those profiling sentences for training, and made comparisons. Therefore, there are four models to compare: concat (correct), concat (correct+possibly correct), profile-wise (correct), and profile-wise (correct+possibly correct).

Table 4
Results of automatic evaluation (*pos_cor* stands for “possibly correct”)

model	BLUE	ROUGE-1	ROUGE-2	ROUGE-L	BERT-Score
concat (*correct*)	33.94	60.63	48.03	52.63	83.64
concat (*correct+pos_cor*)	**36.22**	**65.92**	**52.06**	**57.93**	**85.37**
profile-wise (*correct*)	22.38	45.77	35.44	39.96	78.43
profile-wise (*correct+pos_cor*)	25.38	55.17	40.99	47.93	82.07

Table 5
Results of human evaluation

model	precision	recall	F1
concat (*correct*)	64.05 (98/153)	64.05 (98/153)	64.05
concat (*correct+pos_cor*)	56.88 (91/160)	59.48 (91/153)	58.15
profile-wise (*correct*)	**67.47 (112/166)**	**73.20 (112/153)**	**70.22**
profile-wise (*correct+pos_cor*)	45.28 (96/212)	62.75 (96/153)	52.60
LLM+GOLD (our dataset)	100.00 (147/147)	96.08 (147/153)	98.00

5.2. Metrics

We use not only the same automatic utterance-level metrics as in previous studies, but also human ratings of the profiling sentences. In previous studies [4], profiling sentences generated for each utterance were concatenated for automatic evaluation, but our interest is in the degree to which each profiling sentence generated corresponds to the target utterance. Evaluating for simple concatenation would give the same score to an utterance that produces only one short profiling sentence as to one that produces multiple longer profiling sentences. This means that the more profiling sentences an utterance generates, the lower score is associated with a single profiling sentence, and each profiling sentence cannot be evaluated equally. Therefore, in addition to the evaluation of each target utterance as in previous studies, we also conduct a human evaluation of each profiling sentence to see if there are differences between both types of evaluations.

For automatic evaluation of each utterance, a test set of the created dataset is used. BLEU [12], ROUGE [13] and BERT Score [14] are used as automatic evaluation indicators. However, since these evaluation metrics cannot be applied to target utterances without profiling sentences, only target utterances with profiling sentences are considered.

For the human evaluation at the profiling sentence level, we use 100 target utterances randomly selected from the test set of the created dataset. In this case, the target utterance without a profiling sentence is also evaluated as a single case of data, with the case where no profiling sentence is generated as the correct answer. The generated “profiling sentences” are mapped one-to-one to “target utterance, two preceding utterances” and judged manually as correct or incorrect by annotator. Since our goal is to generate all the profiling sentences that a human would infer from the target utterance, we include profiling sentences that are possibly correct in the correct answer and make a binary decision. Three annotators are hired for the human evaluation, and the decision is made by majority vote. We also create a set of profiling sentences ΔP_{t_gold}, all of which are manually extracted for the 100 test sets used in the manual evaluation. For ΔP_{t_gold}, three annotators are asked to write out profiling sentences, and one of the authors checks for duplicates and removes them.

5.3. Experimental results

The experimental results of the automatic evaluation are shown in Table 4. Comparing the combined method and our one-to-one training method, the combined method scored higher overall in the automatic evaluation. This indicates that training and outputting the combined profiling sentences is advantageous for automatic evaluation against the combined correct sentences. Next, when comparing the model trained on only correct answers with the model trained on both correct and possibly correct answers, the model trained solely on correct answers obtained a lower score. This may be because the correct sentence, referred to as the correct answer in the automatic evaluation, included profiling sentences that could potentially be correct in the training. This likely favored the model with a more diverse output.

The experimental results of the human evaluation are shown in Table 5. Here, LLM+GOLD is the result of the dataset itself. The total number of ΔP_{t_gold} is used to calculate the recall and F1 scores. Overall, the results were different from those of the automatic evaluation. In particular, profile-wise (correct), which had the lowest score in the automatic evaluation, resulted in the highest F1 score in the human evaluation. This suggests that automatic evaluation at the utterance level is less correlated with human evaluation at the profiling sentence level.

5.4. Context influence analysis

The profiling sentences generated by the human evaluation are further analyzed to check whether the model trained on the created dataset is able to generate profiling sentences while being influenced by the context. First, for comparison, we manually count the profiling sentences that could not be predicted without context in the ΔP_{t_gold}, a set of profiling sentences that are all extracted manually. The results show that 20.92% (32/153) of the profiling sentences required context. Thus, we see that there are a certain number of profiling sentences in the test set that require context. Next, we analyze the profiling sentences included in the dataset we have created. The results show that 3.40% (5/147) of the profiling sentences referred to the context. This confirms that the proportion of contextual references in LLM profile writing is much lower than that in human profile writing. This is a limitation of using LLMs to extend the dataset, but could be improved as the capabilities of the LLMs increase. The percentage of profiling sentences generated by the learned model is as follows:

- concat (correct): 5.06% (4/79)
- concat (correct + possibly correct): 7.14% (6/84)
- profile-wise (correct): 3.95% (3/76)
- profile-wise (correct + possibly correct): 7.14% (6/84)

We see that the profiling sentence generation model is able to generate profiling sentences with reference to the context at about the same rate as the dataset, although less than the human profiling sentences. Comparing the model trained with only the correct answer and the model trained with both the correct answer and the possibility of the correct answer, it is found that the model trained with both the correct answer and the possibility of the correct answer is more context-referenced. On the other hand, as can be seen from Table 5, this model has lower prediction accuracy, hence achieving both is an issue to be solved in the future.

6. Conclusions and future work

In this study, we proposed a task to generate profiling sentences from target utterances by adding two utterances as context, and created and evaluated a dataset linking profiling sentences that can be inferred from a single utterance. For the generation of profiling sentences, we proposed a one-to-one training method in which profiling sentences are learned from target utterances, in addition to the conventional method of training profiling sentences that can be inferred from a single utterance by concatenating them. For the evaluation, automatic evaluation at the utterance level and human evaluation at the profiling sentence level were performed. The experimental results show distinct discrepancy between automated and human evaluations for inferring profiling sentences. We also confirmed that the proposed training method is as effective as, or even more effective than, existing approaches. Analysis of the generated profiling sentences confirmed that the profiling sentence generation model is able to generate implicit profiling sentences at approximately the same rate as the dataset. In the future, we plan to work on creating a dataset that includes more contextually referenced profiling sentences. We also intend to analyze the relationship between profiling sentences for each interlocutor by converting the utterances in the created dataset into dialogue-by-dialogue data.

7. Limitations

We conducted our experiments in Japanese, but further verification is needed because the results may change if the experiment is conducted in English or any other language. This paper has shown the differences between the automatic evaluation methods used in existing research and human evaluation, but it is very costly. It is important to study more context-based one-to-many automatic evaluation metrics. Although this research focuses only on the chat domain, the method itself may be applicable to other domains such as opinion extraction. Experiments in other domains is desired in the future. Our research is based on dialogue data created by crowdworkers who pretended to be non-existent people. Therefore, our dataset also does not include profiles of real people. However, when considering actual applications, collecting profiles of existing interlocutors may be a problem from the perspective of privacy. We do not recommend collecting user information without their permission. To ensure smooth and reliable interactions between users and systems, sentence profiling is essential, and this research focuses on achieving that goal.

Acknowledgments

This research was supported by JST Next Generation Challenging Researchers Program JPMJSP2119 and JST CREST Grant Number JPMJCR20D2, Japan.

References

[1] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, 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), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2204–2213. URL: https://aclanthology.org/P18-1205. doi:10.18653/v1/P18-1205.

[2] J. Xu, A. Szlam, J. Weston, Beyond goldfish memory: Long-term open-domain conversation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5180–5197. URL: https://aclanthology.org/2022.acl-long.356. doi:10.18653/v1/2022.acl-long.356.

[3] H. Song, Y. Wang, K. Zhang, W.-N. Zhang, T. Liu, BoB: BERT over BERT for training persona-based dialogue models from limited personalized data, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 167–177. URL: https://aclanthology.org/2021.acl-long.14. doi:10.18653/v1/2021.acl-long.14.

[4] R. Ribeiro, J. P. Carvalho, L. Coheur, PGTask: Introducing the task of profile generation from dialogues, in: S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, M. Alikhani (Eds.), Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue,

Association for Computational Linguistics, Prague, Czechia, 2023, pp. 183–189. URL: https://aclanthology.org/2023.sigdial-1.17. doi:10.18653/v1/2023.sigdial-1.17.

[5] P.-E. Mazaré, S. Humeau, M. Raison, A. Bordes, Training millions of personalized dialogue agents, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2775–2779. URL: https://aclanthology.org/D18-1298. doi:10.18653/v1/D18-1298.

[6] S. Welleck, J. Weston, A. Szlam, K. Cho, Dialogue natural language inference, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3731–3741. URL: https://aclanthology.org/P19-1363. doi:10.18653/v1/P19-1363.

[7] J.-C. Gu, Z. Ling, Y. Wu, Q. Liu, Z. Chen, X. Zhu, Detecting speaker personas from conversational texts, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1126–1136. URL: https://aclanthology.org/2021.emnlp-main.86. doi:10.18653/v1/2021.emnlp-main.86.

[8] M. Shinji, T. Masashi, I. Toshihiko, Verification of LLM's ability to extract speaker information from utterance and context during chats, in: Proceedings of the Thirtieth Annual Meeting of the Association for Natural Language Processing, 2024, pp. 3050–3054.

[9] H. Sugiyama, M. Mizukami, T. Arimoto, H. Narimatsu, Y. Chiba, H. Nakajima, T. Meguro, Empirical analysis of training strategies of transformer-based Japanese chit-chat systems, 2021. arXiv:2109.05217.

[10] S. Bae, D. Kwak, S. Kang, M. Y. Lee, S. Kim, Y. Jeong, H. Kim, S.-W. Lee, W. Park, N. Sung, Keep me updated! memory management in long-term conversations, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 3769–3787. URL: https://aclanthology.org/2022.findings-emnlp.276. doi:10.18653/v1/2022.findings-emnlp.276.

[11] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[12] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[13] C.-Y. Lin, E. Hovy, Manual and automatic evaluation of summaries, in: Proceedings of the ACL-02 Workshop on Automatic Summarization, Association for Computational Linguistics, Phildadelphia, Pennsylvania, USA, 2002, pp. 45–51. URL: https://aclanthology.org/W02-0406. doi:10.3115/1118162.1118168.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, CoRR abs/1904.09675 (2019). URL: http://arxiv.org/abs/1904.09675. arXiv:1904.09675.