General Paper

Collection of Meta Information with User-Generated Question Answer Pairs and its Reflection for Improving Expressibility in Response Generation

Takashi Kodama†, Ryuichiro Higashinaka††, Koh Mitsuda†††, Ryo Masumura†††, Yushi Aono†††, Ryuta Nakamura††††, Noritake Adachi†††† and Hidetoshi Kawabata††††

This paper concerns the problem of realizing consistent personalities in neural conversational modeling by using user generated question-answer pairs as training data. Using the framework of role play-based question-answering, we collected single-turn question-answer pairs for particular characters from online users. Meta information was also collected such as emotion and intimacy related to question-answer pairs. We verified the quality of the collected data and, by subjective evaluation, we also verified their usefulness in training neural conversational models for generating responses reflecting the meta information, especially emotion.

Key Words: Data Collection, Meta Information, User-generated Content, Utterance Generation, Question-Answering

1 Introduction

Much attention has been paid to non-task oriented dialogue systems from their social and entertainment aspects (Wallace 2009; Banchs and Li 2012; Higashinaka et al. 2014). To make such systems more engaging, it is necessary that they exhibit consistent personalities. In neural conversational modeling (Sordoni et al. 2015; Vinyals and Le 2015; Shang et al. 2015; Serban et al. 2016; Adiwardana et al. 2020; Roller et al. 2020), various architectures have been proposed to realize consistent personalities, such as incorporating user IDs (Li et al. 2016), personal attributes (Qian et al. 2018), and profile text (Zhang et al. 2018). However, the problem is that it is still costly to collect dialogue data related to particular personalities for training such neural models.

In this paper, we deal with single question-answer pairs as dialogue data. For efficiently collecting such dialogue data for a particular character, a data collection method called “role play-based question-answering” has been proposed (Higashinaka et al. 2013b, 2018) (See Section 2 for

† Kyoto University, Graduate School of Informatics; this is the work done during the internship at NTT.
†† NTT Corporation; Nagoya University, Graduate School of Informatics
††† NTT Corporation
†††† DWANGO Co., Ltd.
Kodama et al. Collection of Meta Information with User-Generated Question Answer Pairs

details). In this method, fans of a particular character voluntarily provide question-answer pairs by playing the role of that character. It has been demonstrated that high-quality question-answer pairs can be efficiently collected and that dialogue systems that exhibit consistent personalities can be realized (Higashinaka et al. 2018).

This paper extends this work and aims to collect question-answer pairs for particular characters together with other pieces of information (called “meta information”), such as emotion and intimacy levels. The aim of collecting this additional data is to realize dialogue systems whose utterances can be controlled to reflect such meta information (Zhou et al. 2018; Zhou and Wang 2018; Song et al. 2019). This is a useful feature when we want to realize systems that are affective and can become more intimate as an interaction progresses (Zhou and Wang 2018). We verify the quality of the collected data and empirically show that conversational models that exhibit consistent personalities as well as meta information, especially emotion, can be successfully realized by using voluntarily provided user-generated question-answer pairs. Note that, in this paper, we regard personality as the character-ness of a real or fictional character.

In what follows, we first describe the idea of role play-based question-answering followed by our data collection of question-answer pairs and meta information. Then, in Section 3, we describe our approach for training conversational models that take the meta information into account. In Section 4, we describe experiments conducted to verify our approach; we performed both objective and subjective evaluations using the data collected for three characters. Finally, we summarize the paper and mention future work.

2 Data Collection using Role Play-based Question-answering

Role play-based question-answering (Higashinaka et al. 2013b, 2018) is a data collection framework in which multiple users (typically fans) play the role of a certain character and respond to questions by online users (who can also be fans). Since the fans are knowledgeable about the character and find it amusing to answer questions in the role of their favorite character and online users can ask various questions to their favorite character, this framework can motivate users to voluntarily provide dialogue data centering around a particular character. Higashinaka et al. (2018) showed that fans are highly motivated to provide data and that the collected data are of high quality to realize dialogue systems exhibiting consistent personalities.

In this study, we use this framework to collect question-answer pairs together with other pieces of meta information. More specifically, we collect emotion and intimacy levels from fans in addition to question-answer pairs.
2.1 Data collection including meta information

We collected dialogue data (single-turn question-answer pairs) for three famous characters in Japan: Ayase Aragaki (Ayase), Hime Tanaka (Hime), and Hina Suzuki (Hina). Ayase is a fictional character in the novel series “Ore no imouto ga konnani kawaii wake ga nai” (My Little Sister Can’t Be This Cute). Her character is often referred to as a “yandere.” According to Wikipedia, Yandere characters are mentally unstable, incredibly deranged, and use extreme violence or brutality as an outlet for their emotions. Hime and Hina are virtual YouTubers and form a duo called “HIMEHINA.” Hime’s character is friendly, and Hina has a goofy and laid-back character.

Question-answer pairs were collected on the websites established in the characters’ fan communities. Figures 1 and 2 show screenshots of the websites for Ayase and Hime, respectively. The website for Hina is identical to that for Hime except that the images used were those of Hina. Users can ask the characters questions by using a text-field interface, and users who want to play the role of the characters can post answers. Users can post questions and answers at any time; that is, the interaction is asynchronous. Multiple answers can be posted to the same question. In addition, users can input meta information at the same time when posting their answers. The meta data that we collected were of two kinds:

**Emotion** is a label provided for an answer. This indicates the emotion behind the answer, such as angry or happy. The list of emotions is different for each character. There are 10 types of emotion labels for Ayase, including “Normal,” “Stumped,” and “Angry.” The labels were decided on the basis of emotions that she exhibits in the novel series in which she appears. For Hime and Hina, we had slightly different emotion labels decided on the basis of their behavior on YouTube. We employed the notion of basic emotions for ease of annotation for online users (Ekman 1992).

**Intimacy** is a label provided for an answer. It indicates how close each respondent is feeling to the questioner, and its value is discrete, taking one integer from 1 (least intimate; intimacy level for a stranger) to 5 (most intimate; intimacy level for a family member).

Since the intimacy feature had not been developed when collecting data for Ayase, we collected only emotion labels for Ayase and collected both kinds of meta information for Hime and Hina. Table 1 shows the collected data examples.

2.2 Statistics of collected data

The statistics of collected question-answer pairs are shown in Table 2. There were 15,179, 12,746, and 10,739 pairs collected for Ayase, Hime, and Hina, respectively. They were collected
within a period shorter than one month, indicating that the framework is efficient for collecting dialogue data. Note that the users who provided data were not paid for their effort; the work was totally voluntary.

As for the meta information, distributions of emotion labels for Ayase are shown in Figure 3. Those for Hime and Hina are shown in Figure 4. It can be seen that the emotion labels for Ayase are rather equally distributed, whereas “Joyful” seems dominant for Hime and Hina, representing their personality. The distributions of intimacy labels for Hime and Hina are shown in Figure 5. For both characters, there were few “1” labels. The dominant labels were “3” and “5.”
Table 1 Examples of collected data. Numbers in “Meta” column indicate intimacy level. The questions and answers are originally in Japanese and were translated by the authors.

| Data | Question | Answer | Meta |
|------|----------|--------|------|
| Ayase | Ayase-tan! What are your plans for the holidays? | I don’t want to tell you! | Angry |
|       | Happy birthday to Ayase-tan!! | Today is not my birthday... | Stumped |
| Hime  | What is your favorite food? | Of course, steamed meat buns!! | Joyful,3 |
|       | Do you blush when you’re told you’re cute? | I might be surprised if you say that suddenly! | Surprised,4 |
|       | What hairstyle do you want to get? | I want to get the same hairstyle as Hina!! | Joyful,4 |
|       | What do you want now? | Hmm... A new umbrella!! | Joyful,2 |
| Hina  | Do you like Hime!?! | I like her! | Favorable,5 |
|       | The moon is beautiful. | That’s right. | Normal,1 |
|       | What made you cry recently? | The other day, I cried while watching an anime. | Sad,5 |
|       | Do you like tea with milk? | Yes! I love it!! | Joyful,5 |

Table 2 Statistics of collected question-answer pairs for Ayase, Hime, and Hina.

|                        | Ayase | Hime | Hina |
|------------------------|-------|------|------|
| # of question-answer pairs | 15,179 | 12,746 | 10,739 |
| # of questions         | 15,179 | 12,746 | 10,739 |
| # of unique questions  | 6,636  | 5,982 | 5,148 |
| Avg # of tokens per question | 12.3  | 11.3  | 11.4  |
| Avg # of letters per question | 20.7  | 18.2  | 18.3  |
| # of unique tokens in questions | 6,514 | 5,761 | 5,348 |
| # of answers           | 15,179 | 12,746 | 10,739 |
| # of unique answers    | 14,587 | 12,420 | 10,574 |
| Avg # of tokens per answer | 13.9  | 14.6  | 14.1  |
| Avg # of letters per answer | 24.4  | 22.2  | 22.1  |
| # of unique tokens in answers | 8,864 | 6,745 | 7,500 |

2.3 Data quality

2.3.1 Procedure

To confirm the quality of the collected question-answer pairs with online users, we conducted a subjective evaluation by using human judges who do not include the authors. We asked potential judges to answer a questionnaire on a five-point Likert scale (1: not knowledgeable, 5: very knowledgeable) for their knowledge levels about each character, and only used those whose self-declared knowledge level was three or more. As a result, eleven, nine, and nine judges participated in the evaluation of question-answer pairs for Ayase, Hime, and Hina, respectively.
Kodama et al. Collection of Meta Information with User-Generated Question Answer Pairs

![Distribution of emotion labels for Ayase](image1)

![Distributions of emotion labels for Hime and Hina](image2)

![Distributions of intimacy labels Hime and Hina](image3)

The judges rated each answer by their degree of agreement to the following statements on a five-point Likert scale (1: strongly disagree, 5: strongly agree).

**Naturalness**: The answer is appropriate for the character’s response.

**Reflection**: The answer reflects the meta information (emotion or intimacy).

When judging the naturalness, the judges were shown pairs of a question and a user-generated answer. 100 unique question-answer pairs were randomly selected from the collected question-answer pairs for this evaluation.

When judging the reflection of meta information, the judges were shown a tuple of a question, the meta information, and the user-generated answer. 100 unique tuples were randomly selected from the collected data for this evaluation. As a control, we prepared 100 unique tuples with meta information randomly replaced with different meta information.

### 2.4 Results

Table 3 shows the evaluation results. In terms of naturalness, all three characters attained...
high scores. This shows that even though role play-based question-answering does not pay users for their efforts, it can be used to collect appropriate responses for characters. This conforms to the results shown in (Higashinaka et al. 2018). Here, we define characteristic responses as those having a naturalness score of 3 or more. Under this definition, the percentages of characteristic responses in the collected data for Ayase, Hime, and Hina were 86%, 99%, and 98%, respectively. This indicates that the responses are of high naturalness.

For the reflection of emotion, when we look at “Actual,” we seem to have good quality emotion labels. When we compare “Actual” vs. “Random” (randomly replaced emotion labels), we have a good amount of drop, meaning that the utterances and emotions are well associated in our data. As for the intimacy, the results were different. Although the “Actual” scores were high, the results for “Random” also exhibited high scores (although with a slight drop), meaning that the utterances and intimacy levels are not as associated when compared with the case of emotion; it may be difficult for humans to accurately recognize the level of intimacy from utterances.

Figure 6 shows the proportion of score differences between judges (i.e., score differences between all judge pairs) for naturalness and reflection of meta information. As for meta information,

|                | Ayase | Hime | Hina |
|----------------|-------|------|------|
| Naturalness    | 3.62  (0.83) | 3.97 (1.00) | 3.99 (0.95) |
| Reflection (emotion) | 3.99 (0.73) | 3.01 (0.93) | 3.89 (0.91) |
| Reflection (intimacy) | N/A   | 3.88 (0.99) | 3.66 (1.13) |

Table 3 Results of human evaluation for collected question-answer pairs. Scores were averaged over all judges. Standard deviations are shown in parentheses.

![Figure 6](image-url) Score differences between judges

142
we only used the results for "Actual" for this analysis. Overall, we can see that most score differences are below 1 or less, meaning that the decisions of the judges are similar to each other.

3 Training Conversational Models that Reflect Meta Information

To test whether it is possible to generate utterances that reflect meta information collected by role-play based question answering, using the collected question-answer pairs as well as meta information, we train neural conversational models. Figure 7 illustrates the overview of our training procedure. Since we want to generate utterances that reflect meta information, we adopt a model architecture that can take such additional information into account in decoding. Below, we explain how we train such conversational models.

3.1 Pre-training with meta information

The amount of data collected for Ayase, Hime, and Hina may be too small to learn a generation model from scratch. Therefore, we decided to pre-train the models with a large amount of data. In this study, we used the large number of general question-answer pairs that we collected previously when developing our question answering system (Higashinaka et al. 2013a). This dataset, "General QA pairs," was collected via crowdsourcing. Crowdworkers were given topics and wrote questions and answers related to the topics; each human intelligence task (HIT) asked each worker to provide 10 question-answer pairs for a topic. There are 500K question-answer pairs in this dataset. For example, for the topic Mt. Fuji, the dataset includes question-answer...
pairs such as Q: “Do you know the height of Mt. Fuji?” and A: “It’s 3,776 meters,” Q: “Is Mt. Fuji the highest mountain in Japan?” and A: “Yes, it is.”

Meta information (i.e., emotion and intimacy labels) is not included in General QA pairs, which may be problematic in the later fine-tuning stage because pre-training without meta information might lead to models that ignore the meta information. Therefore, we trained classifiers for meta information from our data for Ayase, Hime, and Hina and automatically annotated General QA pairs with such meta information. We call the dataset annotated in this way “Auto-annotated general QA pairs.” By performing pre-training using Auto-annotated general QA pairs, a model is likely to take into account the meta information appropriately when fine-tuning with the data with meta information.

3.2 Generation models

Currently, pre-trained language models are showing promising results in a wide variety of natural language processing tasks (Devlin et al. 2019; Radford et al. 2019; Yang et al. 2019; Liu et al. 2020). Such models can more accurately capture the meaning of words depending on the context with a massive amount of training data, enabling them to be applied to fine-tuning for particular downstream tasks. Since BERT (Bidirectional Encoder Representations from Transformers), which is a pre-trained language model, has been found to be useful for generation tasks (Zhang et al. 2019), we use it in our work. Specifically, we adopt the dual-source BERT encoder-decoder model (Junczys-Dowmunt and Grundkiewicz 2018; Correia and Martins 2019). The model allows for the incorporation of additional information. The model was originally proposed for automatic post-editing in machine translation. It takes two inputs: source text in the source language and a tentative machine translation result for that text as additional information. It then outputs target text in the target language. We use this model for our generation models. In this study, as additional information, we used meta information instead of tentative machine translation results.

3.3 Classifiers for meta information

We need classifiers for meta information for creating Auto-annotated general QA pairs. To realize the classifiers, we created BERT-based classifiers with an additional multi-layer perceptron (MLP) layer, using the representation encoded by BERT as input. We used bert-base-multilingual-cased\(^1\) as the BERT model. Tokenization was done by using a SentencePiece\(^2\) (Kudo

\(^1\) https://github.com/google-research/bert
\(^2\) https://github.com/google/sentencepiece
Kodama et al. Collection of Meta Information with User-Generated Question Answer Pairs

and Richardson (2018) model trained with Japanese Wikipedia. The vocabulary size was 32K.

For emotion, there were 10 classes for Ayase and 8 classes for Hime and Hina. In the case of intimacy level, there were five classes for Hime and Hina. For all classifications, we used the cross-entropy as the loss function. We trained the models with a batch size of 32 and a learning rate of 0.00002 for 3 epochs.\(^3\) The classifiers were trained by using the collected data for Ayase, Hime, and Hina and were applied to General QA pairs to create Auto-annotated general QA pairs.

For Ayase, when training the classifiers, we first split the Ayase data randomly into 9/1 for training/test\(^4\) and trained a classifier by using the training set. The accuracy of our classifier was 43.4% using the test set. Compared with the majority baseline that labels all examples as “Normal” and whose accuracy was 30.4%, we considered this accuracy to be reasonable. As for Hime and Hina, we also split the data randomly into 9/1 for training/test. The training procedure was the same as that for Ayase. The classifiers were created for each kind of meta information, that is, emotion and intimacy. The accuracies of the emotion classifiers (majority baseline accuracies in parentheses) for Hime and Hina were 62.1% (54.6%) and 63.7% (59.1%), respectively. The accuracies of the intimacy classifiers for Hime and Hina were 40.5% (33.4%) and 41.2% (36.7%), respectively. Although the accuracies are not that high, we consider them to be reasonable since they outperformed the baselines. The classification accuracies for emotion were much higher than those for intimacy, probably because it is more difficult to distinguish intimacy levels as discussed in Section 2.4.

Tables 4, 5, and 6 show the distributions in automatic annotation for Ayase’s emotion, Hime's and Hina’s emotion, and Hime’s and Hina’s intimacy. Although general QA pairs are mainly annotated with “Normal” for emotion and “3” or “5” for intimacy, there are quite a few kinds of labels appearing in the classifier results; e.g., the proportion of Hina’s “Joyful” was 35.7% (= 176,150/493,348), making the classification results possibly useful as training data for reflecting emotion/intimacy.

|       | Normal | Stumped | Angry | Happy | Scornful | Shy | Sad | Surprised | Doubtful | Sulky |
|-------|--------|---------|-------|-------|----------|-----|-----|-----------|----------|-------|
| Ayase | 429,606| 41,765  | 1,998 | 14,482| 4,959    | 431 | 76  | 30        | 1        | 0     |

Table 4 Distribution of automatically annotated labels for Ayase’s emotion. The numbers denote the number of utterances.

\(^3\) We followed the settings as shown in https://github.com/huggingface/transformers.

\(^4\) Note that the training data here correspond to the training and development data in Section 4.
Table 5 Distribution of automatically annotated labels for Hime’s and Hina’s emotion. The numbers denote the number of utterances.

|       | Joyful | Favorable | Normal | Surprised | Scared | Sad | Angry | Hateful |
|-------|--------|-----------|--------|-----------|--------|-----|-------|---------|
| Hime  | 140,449| 10,185    | 318,207| 380       | 1,403  | 22,695| 29    | 0       |
| Hina  | 176,150| 24,464    | 291,260| 138       | 0      | 1,336 | 0     | 0       |

Table 6 Distribution of automatically annotated labels for Hime’s and Hina’s intimacy. The numbers denote the number of utterances.

|       | 1     | 2    | 3    | 4    | 5    |
|-------|------|------|------|------|------|
| Hime  | 2    | 354  | 434,271| 10  | 58,711|
| Hina  | 0    | 8,032| 130,654| 941 | 353,721|

4 Experiments

4.1 Models for comparison

We trained our conversational models and evaluated their performance. We trained three models for comparison:

- wo-Meta: pre-trained using General QA pairs (without meta information) and fine-tuned without meta information using role play-based QA pairs.
- w-Meta: pre-trained using General QA pairs (without meta information) and fine-tuned with meta information using role play-based QA pairs.
- w-Meta+Anno: pre-trained using Auto-annotated general QA pairs (with meta information) and fine-tuned with meta information using role play-based QA pairs.

We assumed that, by comparing the results for wo-Meta and w-Meta, we could check whether the meta information collected during role play-based question-answering was useful in generating responses that reflect the meta information. By comparing w-Meta and w-Meta-Anno, we could check whether the automatic annotation of meta information was useful for pre-training models. Note that the aim of this paper is to verify whether utterances that reflect meta information can be generated with user-generated question-answer pairs. We used OpenNMT-APE\(^5\) for training the models with default parameters. OpenNMT-APE implements the dual-source BERT encoder-decoder model that allows for the incorporation of additional information. We used the same BERT and sentencepiece models as in Section 3.3. The collected data were randomly split into train, development, and test sets with the ratio of 8/1/1 in this experiment.

\(^5\) https://github.com/deep-spin/OpenNMT-APE
4.2 Automatic evaluation

Tables 7 and 8 show the results of the automatic evaluation against the test data for emotion and intimacy, respectively. We used perplexity, distinct-1,2, and BLEU-1,2,3,4 as evaluation metrics (Liu et al. 2016). Perplexity measures the adequacy of language models. Distinct metrics measure the diversity of expressions in generated utterances, and BLEU metrics measure the accuracy of generated utterances in terms of lexical overlaps with references.

From the tables, we can see that, for emotion, w-Meta+Anno seems to have achieved good results for Ayase and Hina, although the results were mixed for Hime. Across the characters, it can be seen that w-Meta+Anno performed the best for Distinct-2, indicating that our pre-training with Auto-annotated general QA pairs had positive effects in terms of improving the variety in the utterances. For intimacy, the tendencies for Hime and Hina were different with no particular models outperforming others except w-Meta+Anno for Distinct-1 and Distinct-2. This is in line with the results for emotion.

| Data  | Method  | Perplexity | Distinct-1 | Distinct-2 | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|-------|---------|------------|------------|------------|--------|--------|--------|--------|
| Ayase | wo-Meta | 10.14      | 12.46      | 25.89      | 6.82   | 6.14   | 5.52   | 4.80   |
|       | w-Meta  | 10.13      | 12.72      | 26.41      | 6.77   | 6.14   | 5.54   | 4.80   |
|       | w-Meta+Anno | 9.99   | 11.99      | 26.95      | 6.99   | 6.34   | 5.72   | 5.00   |
| Hime  | wo-Meta | 13.28      | 13.65      | 33.05      | 9.06   | 8.31   | 7.61   | 6.63   |
|       | w-Meta  | 13.27      | 13.74      | 33.05      | 9.04   | 8.26   | 7.64   | 6.68   |
|       | w-Meta+Anno | 13.22 | 13.31      | 33.14      | 9.05   | 8.27   | 7.53   | 6.54   |
| Hina  | wo-Meta | 15.65      | 13.44      | 33.20      | 7.86   | 7.15   | 6.59   | 5.94   |
|       | w-Meta  | 15.65      | 13.53      | 33.27      | 7.90   | 7.16   | 6.62   | 5.95   |
|       | w-Meta+Anno | 15.69 | 13.77      | 33.66      | 8.12   | 7.35   | 6.76   | 6.07   |

Table 7 Results of automatic evaluation for emotion

| Data  | Method  | Perplexity | Distinct-1 | Distinct-2 | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|-------|---------|------------|------------|------------|--------|--------|--------|--------|
| Hime  | wo-Meta | 13.28      | 13.65      | 33.05      | 9.06   | 8.31   | 7.61   | 6.63   |
|       | w-Meta  | 13.28      | 13.71      | 33.19      | 8.92   | 8.14   | 7.49   | 6.53   |
|       | w-Meta+Anno | 13.25 | 13.93      | 34.84      | 9.38   | 8.46   | 7.70   | 6.76   |
| Hina  | wo-Meta | 15.65      | 13.44      | 33.20      | 7.86   | 7.15   | 6.59   | 5.94   |
|       | w-Meta  | 15.71      | 13.64      | 33.21      | 7.30   | 6.61   | 6.02   | 5.38   |
|       | w-Meta+Anno | 15.83 | 15.14      | 36.33      | 7.81   | 7.04   | 6.40   | 5.70   |

Table 8 Results of automatic evaluation for intimacy
4.3 Human evaluation

4.3.1 Procedure

To assess the quality of the generated responses, we conducted a subjective evaluation. The procedure for this evaluation was the same as that in Section 2.3.1. We used judges whose self-declared knowledge level about the characters was three or more. As a result, eleven, twelve, and twelve judges participated in the evaluation of question-answer pairs for Ayase, Hime, and Hina, respectively. Note that we used a different set of judges from the experiment in Section 2.3.1. The judges rated each output answer by their degree of agreement to the following statements on a five-point Likert scale (1: strongly disagree, 5: strongly agree).

**Naturalness**: The answer is appropriate for the character’s response.

**Reflection**: The answer reflects the meta information (emotion or intimacy).

When judging the naturalness, the judges were shown pairs of an input question and an answer output by each of the generation models. The input to the models was the 100 questions (with meta information when the model requires it) randomly sampled from the test data. The answers for a question to be evaluated were randomly shuffled and presented. We asked the judges to evaluate each output independently and give the same score if the generated response to a question was the same.

When judging the reflection of meta information, the judges were shown a tuple of an input question, the meta information, and the output answer. We prepared two sets of 100 questions as input to the models. One set comprised random samples from the test data. The other set also comprised random samples from the test data, but unseen meta information was used; this set was created by artificially replacing the meta information with different meta information. For instance, when a tuple had an input question (“Do you like Hime!?”) and the meta information (“Favorable”), the meta information was forcibly replaced with another piece of meta information (e.g., “Surprised”), which was randomly selected from labels excluding the original label (“Favorable”). By using this set, it was possible to test the robustness of the models for unseen (possibly discrepant) meta information. We call the former condition “Seen” and the latter “Unseen.”

4.3.2 Results for naturalness

Table 9 shows the results of the human evaluation for naturalness. Among the three models we trained, w-Meta had the best score. However, the differences of the models were small and not statistically significant. Here we used a Wilcoxon rank-sum test as a statistical test with Bonferroni correction. The three models seemed to show the same level of naturalness, which is
good for w-Meta+Anno because this means it can achieve higher distinct scores without a loss of naturalness. The naturalness of generated responses needs to be sufficiently high in order to guarantee the naturalness of utterances that incorporate meta information such as emotion and intimacy. Table 3 shows that, in the collected data, the naturalness of the responses for Ayase, Hime, and Hina were 3.62, 3.97, and 3.99, respectively. Since the data evaluated here were those made by humans, the figures are considered to be the upper bound of the naturalness. The naturalness of the responses generated by the proposed models (w-Meta and w-Meta+Anno) were 3.33, 3.76, and 3.68, respectively, at the lowest, as shown in Table 9. These scores reached 90% or more in relation to the upper bound. This high percentage indicates that the naturalness of the generated responses was sufficiently high.

4.3.3 Results for reflection of emotion

Table 10 shows the results of the human evaluation for the reflection of emotion. We can see that w-Meta+Anno performed the best. For Ayase and Hime, w-Meta+Anno significantly outperformed wo-Meta. For Hina, we did not see a significant difference but the performance was better than wo-Meta. This indicates the effectiveness of our pre-training with Auto-annotated general QA pairs, at least for emotion. When comparing the results of “Seen” and “Unseen,” the scores for “Seen” were generally higher than those for “Unseen,” which is reasonable, indicating

|      | Ayase | Hime | Hina |
|------|-------|------|------|
|      | Emotion | Emotion | Intimacy | Emotion | Intimacy |
| wo-Meta | 3.36 | 3.85 | 3.84 | 3.77 | 3.77 |
| w-Meta | **3.39** | **3.92** | **3.86** | **3.80** | **3.81** |
| w-Meta+Anno | 3.33 | 3.89 | 3.76 | 3.71 | 3.68 |

Table 9 Results of human evaluation for naturalness. Scores were averaged over all judges.

|      | Ayase | Hime | Hina |
|------|-------|------|------|
|      | Seen | Unseen | Seen | Unseen | Seen | Unseen |
| wo-Meta | 3.18 | 2.76 | 3.70 | 2.83 | 3.78 | 2.75 |
| w-Meta | 3.23 | 2.81 | 3.78 | 2.86 | **3.79** | **2.75** |
| w-Meta+Anno | **3.44**** | **3.05**** | **3.86**** | **3.03**** | **3.79** | **2.82** |

Table 10 Results of human evaluation for reflection of emotion. Scores were averaged over all judges. Asterisks (**) indicate whether the value is significantly better than that of “wo-Meta” (p < 0.01). Wilcoxon rank-sum test was used as statistical test with Bonferroni correction.
the difficulty of handling a combination of utterances with unseen emotions.

Tables 11, 12 and 13 show human evaluation scores for each emotion label for Ayase, Hime and Hina, respectively. Here, we used their “Seen” data for analysis. Note that the numbers of samples for each emotion label are different representing the distribution of the original data. As for Ayase, w-Meta+Anno shows a significant improvement in “Angry” and “Happy” over wo-Meta. Concerning Hime, w-Meta+Anno significantly outperformed wo-Meta in “Surprised” and “Angry.” These two characters seem to have achieved their overall improvement by successfully exhibiting several core emotions.

In contrast, there was no particular emotion gaining improvement for Hina. The limited

|                | Normal [25] | Stumped [13] | Angry [16] | Happy [10] | Scornful [8] | Shy [8] | Sad [6] | Surprised [8] | Doubtful [2] | Sulky [4] | Overall [100] |
|----------------|-------------|--------------|-----------|-----------|--------------|--------|-------|--------------|-------------|----------|--------------|
| wo-Meta        | 3.96        | 3.16         | 2.49      | 3.28      | 3.71         | 2.99   | 2.51  | 2.36         | 3.13        | 2.85     | 3.18         |
| w-Meta         | 3.99        | 3.12         | 2.91*     | 3.32      | 3.29         | 2.99   | 2.51  | 2.50         | 3.13        | 2.85     | 3.23         |
| w-Meta+Anno    | 3.84        | 3.37         | 4.21**    | 3.92**    | 3.28         | 3.30   | 1.96†  | 2.09         | 2.58        | 2.73     | 3.44**       |

|                | Joyful [49] | Favorable [21] | Normal [6] | Surprised [8] | Scared [3] | Sad [5] | Angry [7] | Hateful [1] | Overall [100] |
|----------------|-------------|----------------|-----------|--------------|-----------|-------|----------|------------|--------------|
| wo-Meta        | 3.96        | 4.00           | 3.14      | 3.23         | 3.78      | 3.12  | 2.54     | 2.17       | 3.70         |
| w-Meta         | 4.01        | 3.94           | 3.49      | 3.38         | 3.89      | 3.10  | 3.02*    | 2.17       | 3.78         |
| w-Meta+Anno    | 4.00        | 4.01           | 3.50      | 3.98**       | 3.67      | 2.95  | 3.40**   | 3.58**     | 3.86**       |

|                | Joyful [53] | Favorable [15] | Normal [13] | Surprised [9] | Scared [5] | Sad [1] | Angry [1] | Hateful [2] | Overall [100] |
|----------------|-------------|----------------|-------------|--------------|-----------|-------|----------|------------|--------------|
| wo-Meta        | 4.04        | 4.11           | 3.46       | 3.44         | 2.50      | 3.33  | 2.08     | 2.04       | 3.78         |
| w-Meta         | 4.07        | 4.08           | 3.46       | 3.56         | 2.50      | 2.92  | 2.08     | 2.04       | 3.79         |
| w-Meta+Anno    | 4.04        | 4.02           | 3.54       | 3.60         | 2.50      | 3.00  | 2.17     | 2.04       | 3.79         |
Kodama et al. Collection of Meta Information with User-Generated Question Answer Pairs

Fig. 8 Percentages of unique tokens for each emotion. “Gold” indicates the gold responses for test data and “Generated” indicates the generated responses for test data.

performance of Hina is presumably because Hina had few characteristic phrases to learn, whereas Ayase and Hime had some typical phrases related to specific emotions (e.g., “I will call the police!” (“Angry” for Ayase) and “W-What is that!?“ (“Surprised” for Hime)). Figure 8 shows the percentages of unique tokens found in the responses for each emotion label. With regards to “Angry” of Ayase and “Surprised” for Hime, the percentage of “Generated” is much lower than that of “Gold”. This suggests that the models learned typical phrases for these emotions. In contrast, Hina tends to have smaller drops on all emotion labels, suggesting that the model have not obtained the typical expression of each emotion.

4.3.4 Results for reflection of intimacy

Table 14 shows the results of the human evaluation for the reflection of intimacy. We did not observe w-Meta and w-Meta+Anno to be superior. It can be observed that w-Meta is almost the same as wo-Meta and w-Meta+Anno is slightly worse than wo-Meta. Tables 15 and 16 show human evaluation scores for each intimacy label for Hime and Hina, respectively. Here, we used their “Seen” data for analysis. The differences of the models were small and not statistically significant for any intimacy label for both characters.
Figure 9 shows the percentages of unique tokens for each intimacy label. Both characters showed a tendency to have small drops in all intimacy labels. It is assumed that their generation models could not learn typical expressions associated with each intimacy label. Another reason may be that it is difficult for humans to correctly recognize the level of intimacy from utterances alone, as discussed in Section 2.4.

|               | Hime        | Hina        |
|---------------|-------------|-------------|
|               | Seen Unseen | Seen Unseen |
| wo-Meta       | 3.76 3.58   | 3.82 3.47   |
| w-Meta        | **3.78 3.58** | 3.80 3.42   |
| w-Meta+Anno   | 3.77 3.57   | 3.75 3.33   |

Table 14 Results of human evaluation for reflection of intimacy. Scores were averaged over all judges. Daggers (†) indicate whether the value is significantly worse than that of “wo-Meta” (p < 0.05). Wilcoxon rank-sum test was used as statistical test with Bonferroni correction.

|               | 1  | 2  | 3  | 4  | 5  | Overall |
|---------------|----|----|----|----|----|---------|
|               | [8]| [20]| [30]| [17]| [25]| [100]   |
| wo-Meta       | 2.96 | 3.24 | 3.90 | 3.92 | 4.18 | 3.76   |
| w-Meta        | 2.94 | **3.30** | 3.90 | **3.93** | **4.21** | **3.78** |
| w-Meta+Anno   | 2.88 | 3.29 | 4.01 | 3.84 | 4.10 | 3.77   |

Table 15 Results of human evaluation for reflection of each intimacy label for Hime. Scores were averaged over all judges. The numbers in square brackets indicate the number of samples for each intimacy label in the test data. No significant differences were found between the models.

|               | 1  | 2  | 3  | 4  | 5  | Overall |
|---------------|----|----|----|----|----|---------|
|               | [3]| [13]| [26]| [21]| [27]| [100]   |
| wo-Meta       | 3.42 | 3.19 | **3.70** | 4.10 | 4.01 | **3.82** |
| w-Meta        | **3.44** | 3.10 | 3.67 | 4.03 | **4.04** | 3.80   |
| w-Meta+Anno   | 3.39 | **3.23** | 3.61 | 3.94 | 3.95 | 3.75   |

Table 16 Results of human evaluation for reflection of each intimacy label for Hina. Scores were averaged over all judges. The numbers in square brackets indicate the number of samples for each intimacy label in the test data. No significant differences were found between the models.
5 Summary and Future Work

The purpose of this study was to verify whether a natural utterance can be realized for a character reflecting meta information (emotion and intimacy) from question-answer pairs and meta information obtained by role play-based question-answering. We performed experiments using the data of three famous characters in Japan. When training the generation models, we proposed the use of pre-training with data automatically annotated with meta information, which generally led to better performance.

As for emotion, subjective evaluation results indicate that utterances reflecting meta information can be generated although some differences were found depending on the characters; it seems difficult for a character who lacks typical expressions to generate appropriate responses reflecting emotion. We also found that it was difficult to generate responses that reflect intimacy. Although we require future investigation, we attribute this to the lack of typical phrases related to intimacy and the way humans recognize intimacy in utterances.

For future work, we want to improve the quality of our generation models since there seems to be much room for improvement when compared to human performance. It may also be interesting to apply other pre-training methods (Yang et al. 2019; Liu et al. 2020) as well as to incorporate knowledge of the characters in question (Ghazvininejad et al. 2018) in order to enhance the character-ness of the generated utterances. We also want to examine the relationship between the naturalness of a generated response and the degree to which the meta information can be reflected. Although we work on the character-ness of a real or fictional character, we also want to apply our method to personality in general in the future. It is of our particular interests to investigate whether other kinds of meta-information can be collected and used for response
Realizing workable dialogue agents based on our work is one of our next steps, which would require the consideration of emotion/intimacy over multiple turns.

Finally, it should be noted that our proposed method has the risk of being used without the approval of copyright holders. For the appropriate development and application of the proposed method, appropriate procedures for using the characters need to be taken in order to avoid legal and ethical problems.

Acknowledgement

We thank the developers of DWANGO Co., Ltd. for creating the role play-based question-answering Web sites. We also thank the users who participated in role play-based question-answering. This paper is an extended version of our earlier report (Kodama et al. 2020). We would like to thank the Journal of Natural Language Processing for English writing check for this article. We also thank the anonymous reviewers for their insightful comments.

References

Adiwardana, D. D. F., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V. (2020). “Towards a Human-like Open-Domain Chatbot.” ArXiv, abs/2001.09977.

Banchs, R. E. and Li, H. (2012). “IRIS: A Chat-oriented Dialogue System based on the Vector Space Model.” In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 37–42.

Correia, G. M. and Martins, A. F. T. (2019). “A Simple and Effective Approach to Automatic Post-Editing with Transfer Learning.” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3050–3056.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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), pp. 4171–4186.

Ekman, P. (1992). “An Argument for Basic Emotions.” Cognition & Emotion, 6 (3-4), pp. 169–200.

Ghazvininejad, M., Brockett, C., Chang, M.-W., Dolan, B., Gao, J., Yih, W.-t., and Galley, M. (2018). “A Knowledge-grounded Neural Conversation Model.” In Proceedings of the 32nd
Kodama et al. Collection of Meta Information with User-Generated Question Answer Pairs

*AAAI Conference on Artificial Intelligence*, pp. 5110–5117.

Higashinaka, R., Sadamitsu, K., Saito, K., and Kobayashi, N. (2013a). “Question Answering Technology for Pinpointing Answers to A Wide Range of Questions.” *NTT Technical Review*, 11(7).

Higashinaka, R., Dohsaka, K., and Isozaki, H. (2013b). “Using Role Play for Collecting Question-answer Pairs for Dialogue Agents.” In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pp. 1097–1100.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (2014). “Towards an Open-domain Conversational System Fully Based on Natural Language Processing.” In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 928–939.

Higashinaka, R., Mizukami, M., Kawabata, H., Yamaguchi, E., Adachi, N., and Tomita, J. (2018). “Role Play-based Question-answering by Real Users for Building Chatbots with Consistent Personalities.” In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 264–272.

Junczys-Dowmunt, M. and Grundkiewicz, R. (2018). “MS-UEdin Submission to the WMT2018 APE Shared Task: Dual-Source Transformer for Automatic Post-Editing.” In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers*, pp. 822–826.

Kodama, T., Higashinaka, R., Mitsuda, K., Masumura, R., Aono, Y., Nakamura, R., Adachi, N., and Kawabata, H. (2020). “Generating Responses that Reflect Meta Information in User-Generated Question Answer Pairs.” In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 5433–5441. European Language Resources Association.

Kudo, T. and Richardson, J. (2018). “SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing.” In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71.

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., and Dolan, B. (2016). “A Persona-Based Neural Conversation Model.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003.

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016). “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation.” In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer,
L. (2020). “Multilingual Denoising Pre-training for Neural Machine Translation.” ArXiv, abs/2001.08210.

Qian, Q., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018). “Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation.” In Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 4279–4285.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). “Language Models are Unsupervised Multitask Learners.” OpenAI Blog, 1 (8), p. 9.

Roller, S., Dinan, E., Goyal, N., Ju, D. Y., Williamson, M. F., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., and Weston, J. (2020). “Recipes for Building an Open-domain Chatbot.” ArXiv, abs/2004.13637.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). “Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models.” In Proceedings of the 30th AAAI Conference on Artificial Intelligence, pp. 3776–3783.

Shang, L., Lu, Z., and Li, H. (2015). “Neural Responding Machine for Short-Text Conversation.” In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1577–1586.

Song, Z., Zheng, X., Liu, L., Xu, M., and Huang, X. (2019). “Generating Responses with a Specific Emotion in Dialog.” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3685–3695.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015). “A Neural Network Approach to Context-Sensitive Generation of Conversational Responses.” In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 196–205.

Vinyals, O. and Le, Q. (2015). “A Neural Conversational Model.” arXiv, abs/1506.05869.

Wallace, R. S. (2009). The Anatomy of A.L.I.C.E.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). “XLNet: Generalized Autoregressive Pretraining for Language Understanding.” In Advances in Neural Information Processing Systems 32, pp. 5753–5763.

Zhang, H., Cai, J., Xu, J., and Wang, J. (2019). “Pretraining-Based Natural Language Generation for Text Summarization.” In Proceedings of the 23rd Conference on Computational Natural Language Learning, pp. 789–797.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). “Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too?” In Proceedings of the 56th
Kodama et al. Collection of Meta Information with User-Generated Question Answer Pairs

*Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213.

Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018). “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory.” In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 730–738.

Zhou, X. and Wang, W. Y. (2018). “MojiTalk: Generating Emotional Responses at Scale.” In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1128–1137.

**Takashi Kodama:** Takashi Kodama received his M.S. degree from the Graduate School of Informatics, Kyoto University in 2020. He is currently a Ph.D. student at the Graduate School of Informatics, Kyoto University. His research interests are focused on natural language processing and dialogue systems. He is a member of the Association for Natural Language Processing (ANLP).

**Ryuichiro Higashinaka:** Ryuichiro Higashinaka received his B.A. degree in Environmental Information, Masters in Media and Governance, and Ph.D. from Keio University, Kanagawa, in 1999, 2001, and 2008, respectively. He is currently a professor at Nagoya University. He is a visiting distinguished senior researcher at NTT Media Intelligence Laboratories. His research interests include building question-answering systems and spoken-dialogue systems. From November 2004 to March 2006, he was a visiting researcher at the University of Sheffield, UK. He received the “Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology” in 2016. He is a member of the ANLP, Japanese Society for Artificial Intelligence (JSAI), the Information Processing Society of Japan (IPSJ) and the Institute of Electronics, Information and Communication Engineers (IEICE).

**Koh Mitsuda:** Koh Mitsuda received his M.S. degree in Engineering from Tokyo Institute of Technology and joined the Nippon Telegraph and Telephone (NTT) Corporation in 2015. He is currently a researcher at NTT Media Intelligence Laboratories and a Ph.D. student at Tsukuba University. He is presently engaged in research on natural language processing and dialogue systems. He is a member of the ANLP and JSAI.

**Ryo Masumura:** Ryo Masumura received his B.E., M.E., and Ph.D. degrees in
engineering from Tohoku University, Sendai, Japan, in 2009, 2011, and 2016, respectively. Since joining NTT Corporation in 2011, he has been engaged in research on speech recognition, spoken language processing, and natural language processing. He received the Student Award and the Awaya Kiyoshi Science Promotion Award from the Acoustic Society of Japan (ASJ) in 2011 and 2013, respectively; the Sendai Section Student Awards and the Best Paper Prize from the Institute of Electrical and Electronics Engineers (IEEE) in 2011; the Yamashita SIG Research Award and the SIG-NL Excellent Paper Award from the IPSJ in 2014 and 2018, respectively; the Young Researcher Award and the Paper Award from the ANLP in 2015 and 2020, respectively; and the Information and Systems Society (ISS) Young Researcher’s Award in Speech Field and the ISS Excellent Paper Award from the IEICE in 2015 and 2018, respectively. He is a member of the ASJ, IPSJ, ANLP, IEEE, and the International Speech Communication Association (ISCA).

Yushi Aono: Yushi Aono received his Ph.D. degree in engineering science from Osaka University and joined the NTT Corporation in 1999. After working for NTT Cyber Space Laboratories, NTT Advanced Technology, and the NTT Research and Planning Division, he began working as a project manager (principal researcher) in the Cognitive Information Processing Laboratory at NTT Media Intelligence Laboratories. He is engaged in research on speech recognition and speech synthesis and is a member of the IEICE.

Ryuta Nakamura: Ryuta Nakamura joined the Sony Global Manufacturing & Operations Corporation in 2008. As a software engineer, he developed game consoles and life electronics devices. He joined Gatebox Inc. in 2016 as a software engineer. He was in charge of marketing management in 2017. He joined Dwango Co., Ltd. in 2018 as a new business in charge and joined the “Narikiri AI” project. He currently works as a new business manager since 2020. He has been a member of the Kadokawa Digital Strategy Bureau since 2020.

Noritake Adachi: Noritake Adachi joined Nippon Timeshare Co., Ltd. in 2000. He subsequently joined Dwango Co., Ltd. He was a section manager at a smartphone music site. He is currently engaged as a project manager in a conversation AI creation project called the “Narikiri AI” project (2016–present), he was involved in Artificial Intelligence Fundraising (2018), and he is a winner
Kodama et al. Collection of Meta Information with User-Generated Question Answer Pairs of the Horse Racing Programming Competition “Denno Prize” (2015–2018).

Hidetoshi Kawabata: Hidetoshi Kawabata joined Rakuten, Inc. in 2007. He was a Manager for the West Japan Area for Rakuten Market Business till 2011. He joined Dwango Co., Ltd. in 2011. He was in charge of e-commerce business development as a NicoNico Channel sales manager in 2012. He is currently a business manager for the NicoNico Channel and joined the “Narikiri AI” project in 2016. He has also been part of the Kadokawa Digital Strategy Bureau since 2019.

(Received August 1, 2020)
(Revised November 8, 2020)
(Accepted December 12, 2020)