Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5433–5441
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Generating Responses that Reflect Meta Information in User-Generated Question Answer Pairs

Takashi Kodama[1*], Ryuichiro Higashinaka[2], Koh Mitsuda[2], Ryo Masumura[2], Yushi Aono[2], Ryuta Nakamura[3], Noritake Adachi[3] and Hidetoshi Kawabata[3]
[1]Kyoto University, Japan
[2]NTT Corporation, Japan
[3]DWANGO Co., Ltd, Japan
kodama@nlp.ist.i.kyoto-u.ac.jp
{ryuichiro.higashinaka.tp, koh.mitsuda.td, ryou.masumura.ba, yushi.aono.dy}@hco.ntt.co.jp
{ryuuta_nakamura, noritake_adachi, hidetoshi_kawabata}@dwango.co.jp

Abstract

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 utterances reflecting the meta information, especially emotion.

Keywords: 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), 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 for training such neural models.

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

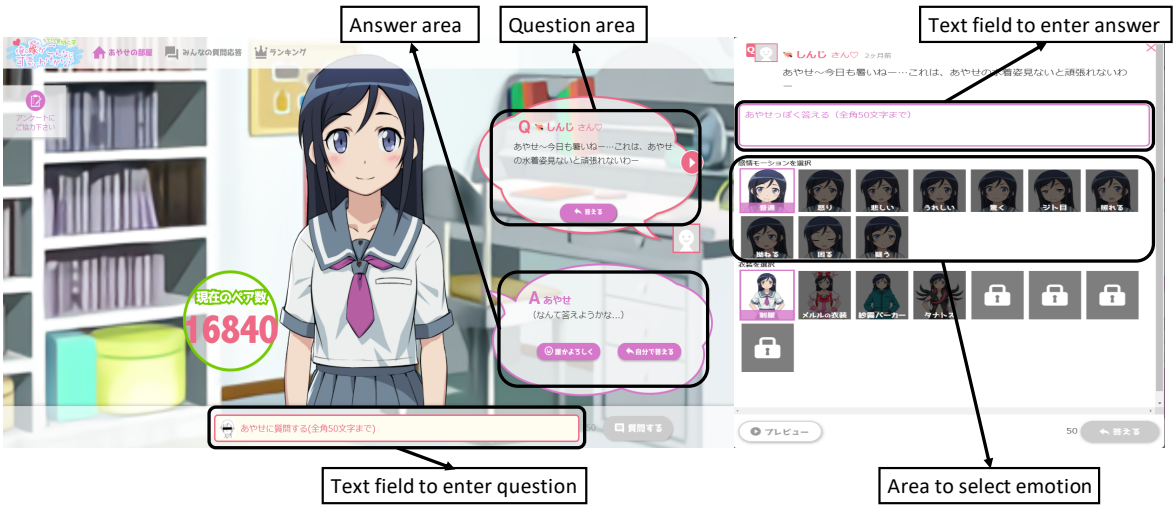
In what follows, we first describe the idea of role play-based question-answering followed by our data collection of question-answer pairs as well as meta information. Then, in Section 3, we describe our approach for training conversational models that take the meta information into account. In Section 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., 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.

*Work carried out during an internship at NTT Corporation.



©Tsukasa Fushimi/ASCII MEDIA WORKS/OIP2 ©BANDAI NAMCO Entertainment Inc. ©2017 Live2D Inc.

Figure 1: Website for Ayase for collecting question-answer pairs

©HIMEHINA Project

Figure 2: Website for Hime for collecting 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 wakeganai" (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 clips of them 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

Table 1: Examples of collected data. Numbers in "Meta" column indicate intimacy level.

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

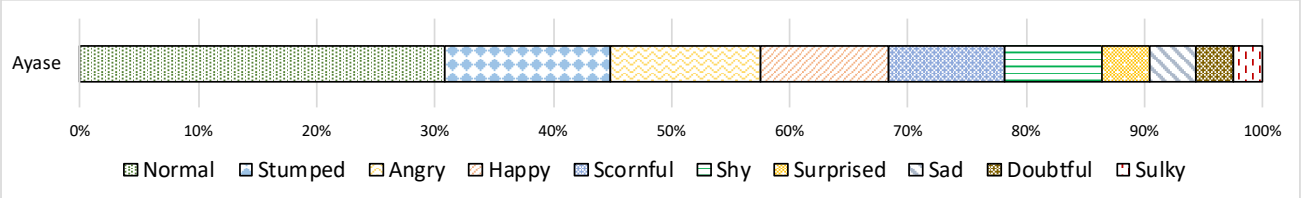


Figure 3: Distribution of emotion labels for Ayase

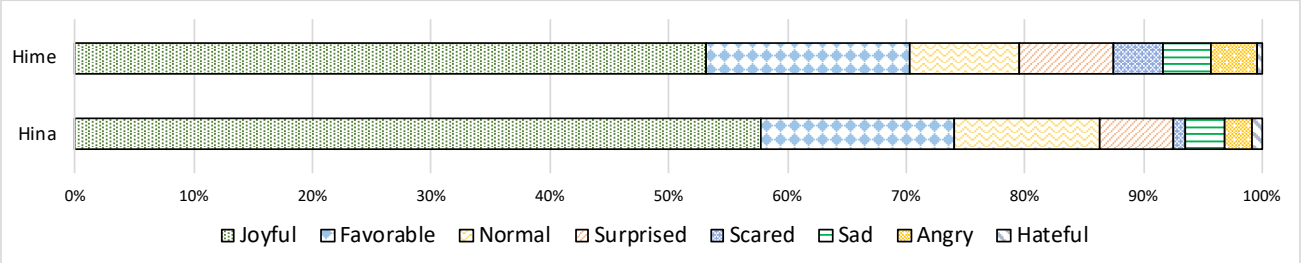


Figure 4: Distributions of emotion labels for Hime and Hina

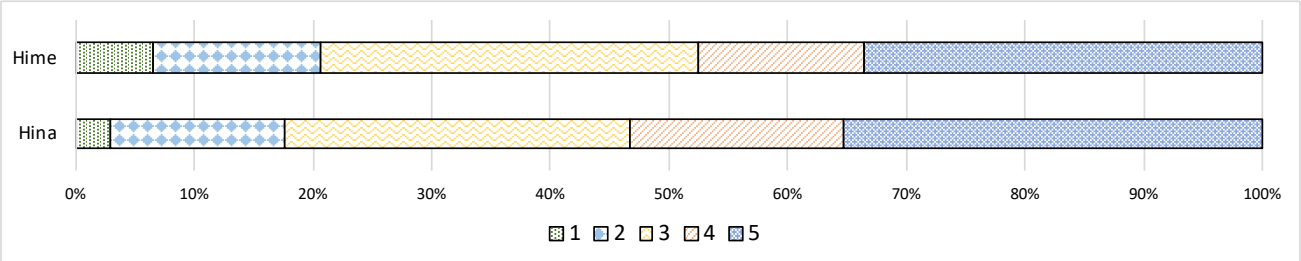


Figure 5: Distributions of intimacy labels Hime and Hina

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	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	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

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 3: Results of human evaluation for collected question-answer pairs. Scores were averaged over all judges.

	Ayase		Hime		Hina	
Naturalness	4.56		4.66		4.76	
	Actual	Random	Actual	Random	Actual	Random
Reflection (emotion)	4.39	3.95	4.57	3.68	4.56	3.73
Reflection (intimacy)	N/A		4.68	4.43	4.65	4.34

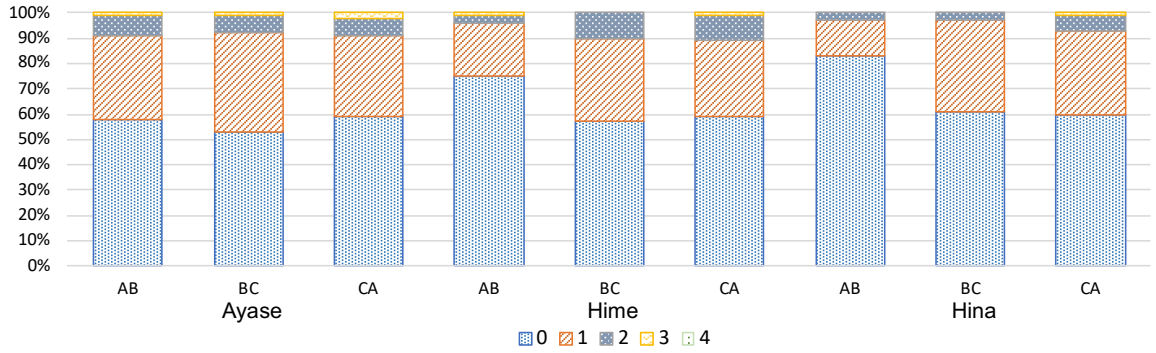


Figure 6: Score differences between each annotator pair for naturalness

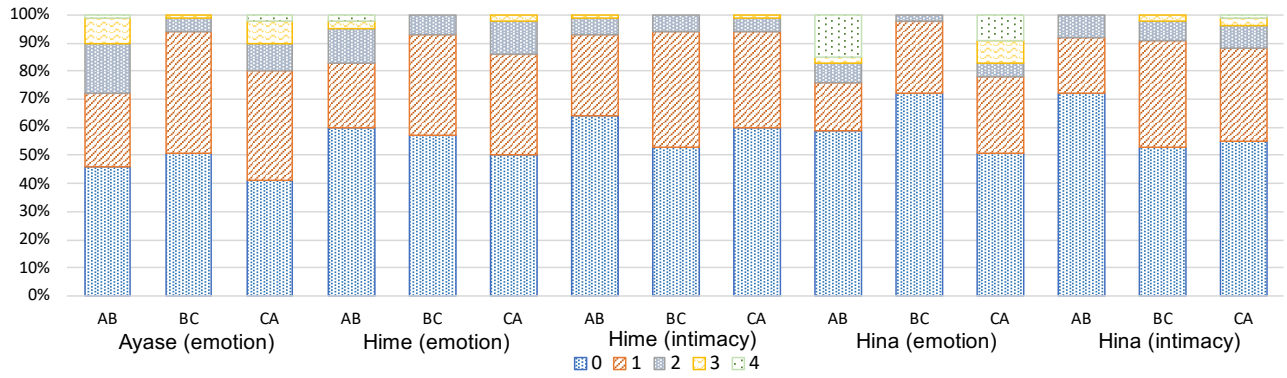


Figure 7: Score differences between each annotator pair for meta information

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. Three judges (not including the authors) evaluated sampled question-answer pairs. Although the number of annotators was small, they are dedicated judges who specialize in text analysis and are knowledgeable about the three characters.

The judges rated each answer by their degree of agreement to the following statements on a five-point Likert scale (1: completely disagree, 5: completely 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 50 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).

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 score differences between each annotator pair for naturalness. Here, our three judges are named A, B, and C; as an example, AB represents a pair of A and B. We can see that most score differences are below 1 or

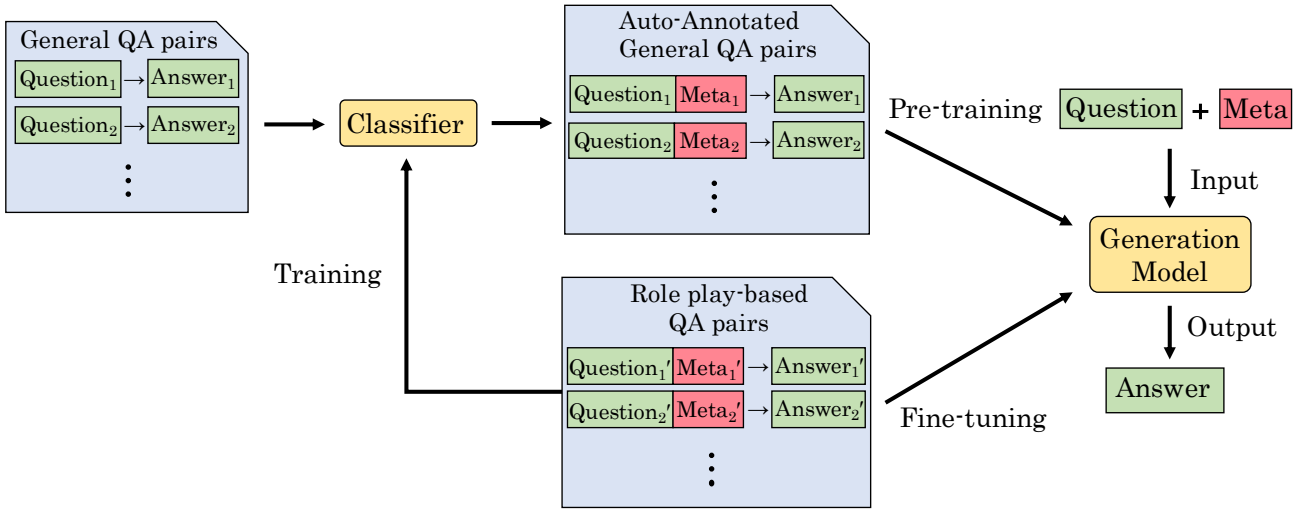


Figure 8: Overview of our approach

less, meaning that the decisions of the judges are similar to each other. Figure 7 shows the score differences for meta information (only the results for "Actual" are used for this analysis). The tendency is similar to that of naturalness, but there seems to be more discrepancy; we observe cases where the difference is three or more in some cases. This indicates that the judgment of emotion/intimacy is more difficult than that of naturalness.

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 8 illustrates an 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., 2013). 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 the 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 by using Auto-Annotated General QA pairs, a model will be able to look at 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. 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 recently been found to be useful for generation tasks (Zhang et al., 2019), we also 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.

For emotion, there were 10 classes for Ayase and 8 classes for Hime and Hina. In the case of intimacy level, there were

Table 4: Results of automatic evaluation 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 5: Results of automatic evaluation for intimacy

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

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.[1] 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[2] 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.

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[3] 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. Tokenization was done by using a SentencePiece[4] (Kudo and Richardson, 2018) model trained with Japanese Wikipedia. The vocabulary size is 32K.

4.2. Automatic evaluation

Tables 4 and 5 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

[1] We followed the settings as shown in `https://github.com/huggingface/transformers`

[2] Note that the training data here correspond to the training and development data in Section 4.

[3] https://github.com/deep-spin/OpenNMT-APE

[4] `https://github.com/google/sentencepiece`

Table 6: Results of human evaluation for naturalness

	Ayase	Hime		Hina	
	Emotion	Emotion	Intimacy	Emotion	Intimacy
wo-Meta	3.88	4.08	4.12	4.07	4.05
w-Meta	3.88	4.06	**4.17**	4.10	**4.08**
w-Meta+Anno	**3.89**	**4.12**	4.01	**4.12**	4.03

Table 7: Results of human evaluation for reflection of meta information. Scores were averaged over all judges. Asterisks (*) indicate whether value of best score is significantly better than "wo-Meta." ** indicates statistical significance p < 0.01, and * indicates p < 0.05. Moreover, + means marginally significant difference (p < 0.10). Wilcoxon rank-sum test was used as statistical test.

	Ayase		Hime				Hina			
	Emotion		Emotion		Intimacy		Emotion		Intimacy	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
wo-Meta	3.66	3.45	4.02	3.65	3.80	3.44	4.07	3.54	**3.97**	3.44
w-Meta	3.66	3.47	4.17**	3.71	**3.82**	**3.50**	4.10	3.54	3.92	**3.46**
w-Meta+Anno	**3.81***	**3.6+**	**4.23***	**3.79**	3.71	3.45	**4.14**	**3.62**	3.74	3.33

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 seem 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.

4.3. Human evaluation

4.3.1. Procedure

To assess the quality of the generated responses, we conducted a subjective evaluation. The procedure was the same as that in Section 2.3.1. The same three judges (not including the authors) evaluated utterances reflecting emotion and intimacy. The judges rated each output answer by their degree of agreement to the following statements on a five-point Likert scale (1: completely disagree, 5: completely 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 6 shows the results of the human evaluation for naturalness. Among the three models we trained, w-Meta+Anno had the best score for emotion, and w-Meta obtained the best score for intimacy. However, the differences of the models were small and not statistically significant. 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.

4.3.3. Results for reflection of meta information

Table 7 shows the results of the human evaluation for the reflection of meta information.

As for emotion, we can see that w-Meta+Anno performed the best. For Ayase and Hime with "Seen" data, w-Meta+Anno significantly outperformed wo-Meta. Although we did not see a statistical significance for Hina, w-Meta+Anno also performed the best. 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," we can see that the improvement of w-Meta+Anno over wo-Meta was limited. The current model does not seem to have the ability to force a generated utterance to exhibit arbitrary emotions. In addition, the scores for "Seen" were generally higher than those for "Unseen," indicating the difficulty of handling a combination of utterances with unseen emotions.

As for intimacy, we did not observe w-Meta-Anno to be superior. Instead, w-Meta seems to be have been better than the others, although there was no statistical difference. The low performance of w-Meta-Anno may be due to the low accuracy of the intimacy classifier. Another reason may be that the judges had difficulty distinguishing the intimacy levels behind utterances. We need further investigation into the cause of w-Meta+Anno not reflecting the assigned intimacy level appropriately.

Overall, since the subjective scores of the models are reasonably high, the results indicate that good-quality utterances that reflect meta information, especially emotion, can be realized with data collected through role play-based question-answering. We also found that pre-training with Auto-Annotated General QA pairs is effective for generating utterances that reflect emotion.

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 from question-answer pairs and meta information obtained by role play-based question-answering. Subjective evaluation results indicate that utterances reflecting meta information can be generated. We confirmed this by utilizing multiple characters and two kinds of meta information (emotion and intimacy). We also showed that the use of pre-training with data automatically annotated with meta information (especially, emotion) is helpful in training generation models. For future work, we want to improve the quality of our generation models. Further investigation is needed in order to realize models that can reflect intimacy. Realizing workable dialogue systems that exhibit emotion and intimacy would also be one of our next steps.

6. Bibliographical references

Banchs, R. E. and Li, H. (2012). IRIS: a chat-oriented dialogue system based on the vector space model. In Proceedings of the ACL 2012 System Demonstrations, pages 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, pages 3050–3056.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Higashinaka, R., Sadamitsu, K., Saito, K., and Kobayashi, N. (2013). Question answering technology for pinpointing answers to a wide range of questions. *NTT Technical Review*, 11(7).

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, pages 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, pages 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 Third Conference on Machine Translation: Shared Task Papers, pages 822–826.

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, pages 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), pages 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, pages 2122–2132.

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 Twenty-Seventh International Joint Conference on Artificial Intelligence, pages 4279–4285.

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 Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, pages 3776–3783. AAAI Press.

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), pages 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, pages 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, pages 196–205.

Vinyals, O. and Le, Q. (2015). A neural conversational model. *ICML Deep Learning Workshop, 2015*.

Wallace, R. S. (2009). The anatomy of a.l.i.c.e. In Robert Epstein, Gary Roberts and Grace Beber (Eds.), *Parsing the Turing Test: Philosophical and Methodological Issues*

in the Quest for the Thinking Computer. pp. 181–210.

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

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 (CoNLL), pages 789–797.

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), pages 1128–1137.

Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In Thirty-Second AAAI Conference on Artificial Intelligence.