Role play-based question-answering by real users for building chatbots with consistent personalities

Ryuichiro Higashinaka[1], **Masahiro Mizukami**[1], **Hidetoshi Kawabata**[2]
Emi Yamaguchi[2], **Noritake Adachi**[2], **and Junji Tomita**[1]
[1]NTT Corporation
[2]DWANGO Co., Ltd.
{higashinaka.ryuichiro, mizukami.masahiro}@lab.ntt.co.jp
{hidetoshi_kawabata, emi_yamaguchi, noritake_adachi}@dwango.co.jp
tomita.junji@lab.ntt.co.jp

Abstract

Having consistent personalities is important for chatbots if we want them to be believable. Typically, many question-answer pairs are prepared by hand for achieving consistent responses; however, the creation of such pairs is costly. In this study, our goal is to collect a large number of question-answer pairs for a particular character by using role play-based question-answering in which multiple users play the roles of certain characters and respond to questions by online users. Focusing on two famous characters, we conducted a large-scale experiment to collect question-answer pairs by using real users. We evaluated the effectiveness of role play-based question-answering and found that, by using our proposed method, the collected pairs lead to good-quality chatbots that exhibit consistent personalities.

1 Introduction

Having a consistent personality is important for chatbots if we want them to be believable (Li et al., 2016; Gordon et al., 2016; Curry and Rieser, 2016; Sugiyama et al., 2017; Akama et al., 2017). Although neural network-based methods are emerging for achieving consistent personalities, their quality is not that high (Li et al., 2016). Therefore, in many systems, question-answer pairs are prepared by hand for consistent responses (Takeuchi et al., 2007; Leuski et al., 2009; Traum et al., 2015). However, the creation of such pairs is costly.

In this study, our aim is to collect a large number of question-answer pairs for a particular character by using role play-based question-answering (Higashinaka et al., 2013a) in which multiple users play the roles of certain characters and respond to questions by online users. The concept is shown in Figure 1. The main idea is that role players collectively represent a single character and that a question is broadcast via a character to all role players. In this way, question-answer pairs can be efficiently collected because there is less burden on people responding, and the entertaining nature of role playing makes people likelier to participate (Ments, 1999). In a small-scale experiment, Higashinaka et al. found that question-answer pairs of a character can be efficiently collected by multiple users and that users are highly motivated to provide questions and answers.

There were two limitations to their work. One was that the experiment was conducted using only a small number of people, who were recruited by the authors. It was not clear if the scheme would work with real users (i.e., users who are not recruited nor paid by researchers). The other limitation was that the applicability of the collected data to the creation of chatbots was not verified. In their small-scale experiment, the maximum number of question-answer pairs for a character was only about 80. This was because users were allowed to register any of their favorite characters, resulting in a small amount of data per character. It was difficult to create a chatbot with such little data.

In this paper, we tackle these limitations by using role play-based question-answering for collecting question-answer pairs from real users. Regarding the second limitation, we limited the characters to two famous ones so as to collect a large number of question-answer pairs per character and create workable chatbots. We conducted a subjective evaluation of the chatbots by using human participants. Our contributions are as follows:

- We verified that role play-based question-

Proceedings of the SIGDIAL 2018 Conference, pages 264–272,
Melbourne, Australia, 12-14 July 2018.

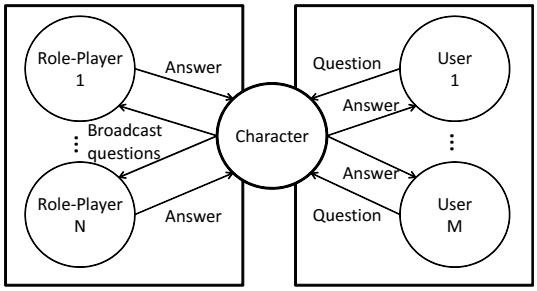


Figure 1: Role play-based question-answering scheme (Higashinaka et al., 2013a).

answering works with real users, collecting a large number of question-answer pairs per character in a short period.

- We proposed a method to create chatbots from collected question-answer pairs and verified that it can lead to good-quality chatbots exhibiting consistent personalities.

We first describe our data collection by using role play-based question-answering with real users. Then, we propose our method for creating chatbots using the collected question-answer pairs. Next, we describe the experiment we conducted to evaluate the quality of the chatbots by using human participants. After covering related work, we summarize the paper and mention future work.

2 Data collection by real users

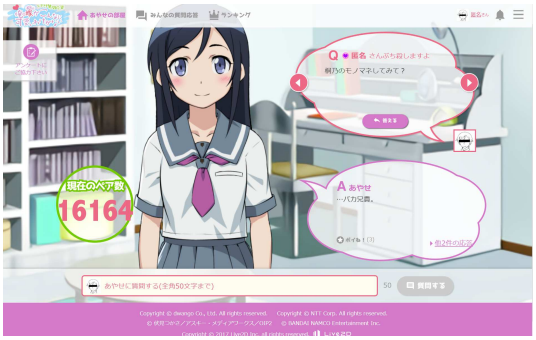
To collect a large number of question-answer pairs per character, we focused on two characters: a real person called Max Murai and a fictional character in a novel, Ayase Aragaki. They are popular characters in Japan and have a large number of fans. We created Web sites in their fan communities so that fans could try role play-based question-answering. We first describe the two characters in more detail and then briefly go over the Web sites. Finally, we present the statistics of the data and look at the results from several aspects.

2.1 Characters

Max Murai His real name is Tomotake Murai (Max Murai is his stage name). Born in 1981, Murai is a CEO of the IT company AppBank but also a YouTuber who specializes in the live coverage of TV games. He is known to have a frank personality.

Ayase Aragaki A fictional character in the novel "Ore no imouto ga konnnai kawaii wakega nai" (My Little Sister Can't Be This Cute), which has sold more than five million copies in Japan in its series. Ayase is not a main character but plays a supporting role. 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.

Figure 2: Web site for Max Murai.



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

Figure 3: Web site for Ayase Aragaki.

2.2 Web sites

On the Japanese streaming service NICONICO Douga[1], each character has a channel for their fans. The channel is limited to subscribers. Through the generosity of this service, we were allowed to establish our Web sites for role play-based question-answering on their channels. Murai has more than 10,000 subscribers; the number of subscribes for Ayase is not disclosed.

We opened the Web sites in March and October 2017 for Murai and Ayase, respectively. Figures 2 and 3 show screenshots of the sites. The appearances of the sites were adjusted to the characters. The users can ask the characters questions by

[1] http://www.nicovideo.jp/

	Murai	Ayase
No. of users who participated	340	333
No. of question-answer pairs	12,959	15,112
No. of questions	7,652	6,482
Average words per question	10.38	13.09
Average letters per question	17.42	20.35
No. of unique words in questions	7,317	6,654
No. of words in questions	79,412	84,838
No. of users who posted questions	284	262
No. of questions per user	22.51	19.47
No. of answers	12,959	15,112
No. of answers per question	1.69	2.33
Average words per answer	7.03	15.27
Average letters per answer	11.59	24.64
No. of unique words in answers	8,666	10,208
No. of words in answers	91,119	230,707
No. of users who posted answers	243	290
No. of answers per user	38.11	45.38

Table 1: Posting statistics.

means of a text-field interface, and users who want to play the role of the characters can post answers. To stimulate interaction, the Web sites show the rankings of users by their number of posts. In addition, a "like" button is placed beside each answer so that when a user thinks the answer sounds very much "like" the character in question, this opinion can be reflected in the number of "likes". The sites were primarily for collecting one-shot question-answer pairs. It was also possible for the Murai site to collect follow-up question-answer pairs, but this function was rarely utilized by users.

2.3 Statistics

The statistics of the postings (at the time of submission) are listed in Table 1. We obtained a total of 12,959 and 15,112 question-answer pairs for Murai and Ayase, respectively. The size of the data is quite large. We want to emphasize that the users were not paid for their participation; they did so voluntarily. This indicates that role play-based question-answering works well with real users. As seen in the table, more than 300 users participated for each character. The questions/answers for Ayase were longer and contained more words and letters.

2.4 Efficiency

Table 2 shows the times when the number of question-answer pairs exceeded certain thresholds. We can see how fast we could collect a few thousand question-answer pairs. For both characters, it took just about a couple of days to reach 2,000 question-answer pairs. For Ayase, the pace was much faster than for Murai, reaching 10,000 question-answer pairs in 18 days. After a certain period, the pace of the postings slowed. Although role play-based question-answering is certainly entertaining, we may need to consider ways to keep users engaged in the interaction. Enabling more sustainable collection of question-answer pairs is future work.

	Murai		Ayase	
Threshold	Hours	Days	Hours	Days
1K	21.36	0.89	25.71	1.07
2K	22.17	0.92	26.88	1.12
5K	1,730.05	72.09	72.21	3.01
10K	2,307.60	96.15	443.73	18.49
12K	2,808.91	117.04	993.37	41.39
15K	N/A	N/A	2,834.26	118.09

Table 2: Time taken to reach certain number of question-answer pairs.

2.5 Quality of the postings

We also evaluated the answers given by the users through subjective evaluation (see GOLD in Tables 4 and 5). We obtained the average naturalness/character-ness scores of around 3.5–4.0 on a five-point Likert scale, indicating that the answers collected through role play-based question-answering were good. However, it was surprising that human users also struggled to obtain scores over 4.0, indicating that generating utterances for a particular character is difficult, even for humans.

2.6 Satisfaction of users

We asked users of the channels to participate in a survey to determine their user satisfaction. We used the same questionnaire as in (Higashinaka et al., 2013a). It consisted of three questions: (Q1) How do you rate the usability of the Web site?, (Q2) Would you be willing to use the Web site again?, and (Q3) Did you enjoy role playing on the Web site? The users answered based on a five-point Likert scale, with one being the lowest score and five the highest. Twenty-three and 36 participants took part in the survey for Murai and Ayase, respectively.

Table 3 shows the results of the questionnaire averaged over all participants. Since these results were obtained from volunteers, they may not reflect the view of all site users. However, the results are encouraging: at the very least, they indicate that there are real users who feel very positively about the experience of role play-based question-answering.

	Questionnaire item	Murai	Ayase
Q1	Usability of Web site	3.74	4.08
Q2	Willingness for future use	4.57	4.56
Q3	Enjoyment of role playing	4.39	4.53

Table 3: Questionnaire results.

3 Creating chatbots from collected question-answer pairs

Now that we have successfully collected a large number of question-answer pairs for our two characters, the next step is to determine if the collected pairs can be useful for creating chatbots that exhibit the personalities of the characters in question; namely, Murai and Ayase. Since the size of the data was not large enough to train neural-generation models (Vinyals and Le, 2015), we opted for a retrieval-based approach in which relevant question-answer pairs are retrieved using an input question as a query and the answer part of the most relevant pair is returned as a chatbot's response. One of the methods we used is a simple application of an off-the-shelf text search engine, and the other is our proposed method, which is more sophisticated and uses neural-translation models for ranking.

3.1 Simple retrieval-based method

This method uses the text search engine LUCENE[2] for retrieval. Questions and answers are first indexed with LUCENE. We use a built-in Japanese analyzer for morphological analysis. Given an input question, the BM25 algorithm (Walker et al., 1997) is used to search for a similar question using the content words of the input question. The answers for the retrieved questions are used as the output of this method. Although simple, this method is quite competitive with other methods when there are many question-answer pairs because it is likely that we will be able to find a similar question by word matching.

3.2 Proposed method

Only using word-matching may not be sufficient. Therefore, we developed a more elaborate method that re-ranks the results retrieved from LUCENE. Our idea comes from cross-lingual question answering (CLQA) (Leuski et al., 2009) and recent advances in neural conversational models (Vinyals and Le, 2015). We also conducted semantic and intent-level matching between questions so that appropriate answer candidates could be ranked higher. Figure 4 shows the flow of this method. Given an input question Q, the method outputs answers in the following steps. The details of some of the key models/modules used in the steps are described later.

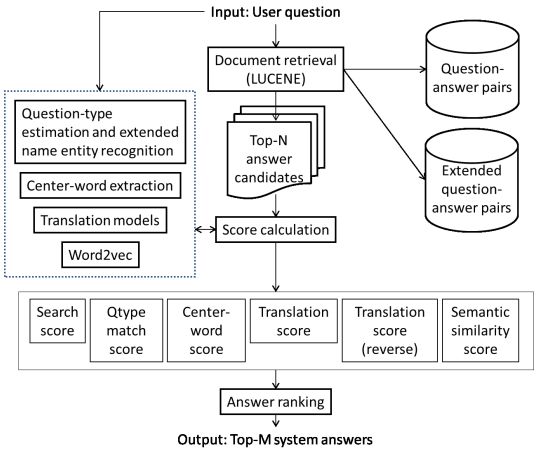


Figure 4: Flow of proposed method.

1. Given Q, LUCENE retrieves top-N question-answer pairs $(Q'_1, A'_1) \ldots (Q'_N, A'_N)$, as described in Section 3.1.
2. The question-type estimation and extended named entity recognition modules estimate the question types of Q and Q' and extract extended named entities (Sekine et al., 2002) contained in A'. The question-type match score is calculated by using the match of the question type and the number of extended named entities in A' requested by Q. See Section 3.3 for details.
3. The center-word extraction module extracts center-words (noun phrases (NPs) that represent foci/topics) from both Q and Q'. The center-word score is 1.0 if one of the center-words of Q is included in those of Q'; otherwise it is 0.0.
4. The translation model is used to calculate the probability that each A' is translated from Q, that is, $p(A'|Q)$. We also calculate the probability bi-directionally, that is, $p(Q|A')$, which has been shown to be effective in CLQA (Leuski et al., 2009). The probabilities are normalized by dividing them by the number of words on the target side. Since the raw probabilities are difficult to integrate with other scores, we sort the question-answer pairs by their probabilities and use their ranks

[2] https://lucene.apache.org/

to obtain the translation scores. That is, if the rank is r, its score is calculated by

$$1.0 - (r - 1)/\text{max\_rank}, \quad (1)$$

where max_rank is the maximum number of elements to be ranked.

5. The semantic similarity model is used to calculate the semantic similarity score between Q and Q'. We use Word2vec (Mikolov et al., 2013) to calculate this score. First, we obtain word vectors (trained from Wikipedia) for each word in Q and Q' and then calculate the cosine similarity between the averaged word vectors.

6. The score calculation module integrates the above scores to obtain a final score:

$$\begin{aligned} &\text{score}(Q, (Q', A')) \\ &= w_1 * \text{search\_score} \\ &+ w_2 * \text{qtypes\_match\_score} \\ &+ w_3 * \text{center-word\_score} \\ &+ w_4 * \text{translation\_score} \\ &+ w_5 * \text{rev\_translation\_score} \\ &+ w_6 * \text{semantic\_similarity\_score} \quad (2) \end{aligned}$$

Here, search_score indicates the score converted from the rank of the search results from LUCENE. The conversion is done using Eq. (1). rev_translation_score indicates the translation score derived from $p(Q|A')$. The $w_1 \ldots w_6$ denote the weights of the scores.

7. The question-answer pairs are sorted by their scores, and top-M answers are returned as output.

3.3 Modules

We describe some of the models/modules used in the above steps.

Question-type estimation and extended named entity recognition We estimated four question types for a question. One is a general question type. We used the taxonomy described in (Higashinaka et al., 2014), which has 16 question subtypes. We trained a logistic-regression based question-type classifier that classifies a question into one of the 16 question types. The other three question types come from an extended named entity taxonomy proposed by Sekine (2002). The taxonomy has three layers ranging from abstract (e.g., Product, Location) to more concrete entities (e.g., Car, Spa, City). We trained a logistic-regression-based classifier that classifies which of the named entity types is requested in a question. We trained a classifier for each layer; thus, we had three classifiers. Using our in-house data, by two-fold cross-validation, the classification accuracies are 86.0%, 84.9%, 76.9%, and 73.5% for the general question type, layer-1, layer-2, and layer-3 question types, respectively. We also extract extended named entities from an answer candidate (A') by using our extended named entity recognizer (Higashinaka et al., 2013b) and check whether the extended named entities corresponding to the layer-1, layer-2, and layer-3 question types of a question (Q) are included in A'.

The qtypes_match_score is calculated as follows: if there is a match of the general question type between Q and Q', the score of one is obtained. Then, the number of extended-named-entity question types covered by the answer candidate is added to this score. Finally, this score is divided by four for normalization.

Center-word extraction We define a center-word as an NP that denotes the topic of a conversation. To extract such NPs from an utterance, we used conditional random fields (CRFs) (Lafferty et al., 2001). For the training and testing, we prepared about 20K sentences with center-word annotation. The sentences were those randomly sampled from our in-house open-domain conversation corpus. The feature template uses words, part-of-speech (POS) tags, and semantic categories of current and neighboring words. The extraction accuracy is 76% in F-measure with our in-house test set.

Translation model We trained a translation model by using a seq2seq model. We trained the model by using the OpenNMT Toolkit[3] with default settings. The translation model learns to translate a question into an answer. By using the trained model, we can obtain the generative probability of an answer given a question; namely $p(A'|Q)$. Since the amount of question-answer pairs was limited, we first trained a model by using our in-house question-answering data comprising 0.5 million pairs. The data were collected using crowd-sourcing. We then adapted the model to our question-answer pairs. The model for $p(Q|A')$ was trained in the same manner by swapping the

[3] http://opennmt.net/

source and target data. To reflect the number of "likes" associated with the answers (see Section 2.2), we augmented the number of samples by their number of "likes"; that is, if a question-answer pair has n "likes", n samples of such a question-answer pair are included in the training data.

3.4 Extending question-answer pairs

When developing our method, we noticed that, in some cases, top-N search results do not contain good candidates because of the lack of question coverage. When the top-N questions do not semantically match reasonably with the input question, the answers are likely to be inappropriate. To have a wider coverage of questions, we extended our question-answer pairs by using Twitter. Our methodology was simple: for each answer A that occurred twice or more in our question-answer pairs, we searched for tweets that resemble A with a Levenshtein distance (normalized by the sentence length) below 0.1. Then, if the tweets had an in-reply-to relationship to other tweets, they were retrieved and coupled with A to form extended question-answer pairs. The reason we focused on an answer that occurred twice or more is mainly due to the efficiency of crawling, but such answers that occur multiple times are likely to be characteristics of the characters in question. We obtained 2,607,658 and 1,032,492 extended question-answer pairs for Murai and Ayase, respectively.

4 Experiments

We conducted a subjective evaluation to determine the quality of chatbots created from our collected question-answer pairs. We first describe how we prepared the data for evaluation and how we recruited participants. We then describe the evaluation criteria. Next, we describe the methods for comparison, in which we compared the methods presented in the previous section with a rule-based baseline and gold data (human-generated data). Finally, we explain the results and present our analyses.

4.1 Data

To create the data for testing, we first randomly split the question-answer pairs into train, development, and test sets with the ratios of 0.8, 0.1, and 0.1, respectively. The splits were made so that the same question would not be included over multiple sets. We used the train and development sets to train the translation models. In addition, the question-answer pairs used by LUCENE for retrieval consisted only of train and development data. For each character, 50 questions were randomly sampled from the test set and used as input questions for this experiment.

4.2 Procedure

We recruited 26 participants each for Murai and Ayase. The participants were recruited mainly from the subscribers of the channels for the two characters. Before taking part in the experiment, they self-declared their levels of knowledge about the characters. Then, they rated the top-1 output of the five methods (shown below) for the 50 questions; they rated at maximum 250 answers (since some methods output duplicate answers, such answers were only rated once). We compensated for their time by giving Amazon gift cards worth about 20 US dollars.

4.3 Evaluation criteria

The participants 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 Not knowing who's speaking, the answer is appropriate to the input question.

Character-ness Knowing that the character in question is speaking, the answer is appropriate to the input question.

The first criterion evaluates the interaction from a general point of view, while the second from the character point of view. Ideally, we want the character-ness to be high, but we want to maintain at least reasonable naturalness when considering the deployment of the chatbots. Note that an utterance can be rated low in terms of naturalness but high in character-ness, or vice-versa: for example, some general utterances, such as greetings, can never be uttered by particular characters.

4.4 Methods for comparison

We compared five methods. A rule-based baseline written in Artificial Intelligence Markup Language (AIML) (Wallace, 2009) was used. The aim of having this baseline is to emulate when we do not have any question-answer pairs available. Although this is a simple rule-based baseline, it is a competitive one because it uses one of the largest rule sets in Japanese.

	All		High		Low	
	Natural	Character	Natural	Character	Natural	Character
(a) AIML	2.93	2.60	2.93	2.49	2.96	2.95
(b) LUCENE	2.80	2.87^{aa}	2.81	2.80^{aa}	2.75	3.10
(c) PROP_WO_EXDB	3.16^{aabb}	3.17^{aabb}	3.17^{abb}	3.09^{aabb}	3.13	$\mathbf{3.42}^{aa}$
(d) PROP	$\mathbf{3.39}^{aabbcc}$	$\mathbf{3.20}^{aabb}$	$\mathbf{3.42}^{aabbcc}$	$\mathbf{3.14}^{aabb}$	$\mathbf{3.32}^{bb}$	3.39^{a}
(e) GOLD	$3.91^{aabbccdd}$	$3.81^{aabbccdd}$	$3.93^{aabbccdd}$	$3.80^{aabbccdd}$	$3.85^{aabbccdd}$	$3.85^{aabbccdd}$

Table 4: Results for Murai. The scores were averaged over the participants. Superscripts indicate whether the value is significantly better than those for the methods denoted with letters; two letters, such as 'aa', indicate statistical significance $p < 0.01$, and a single letter indicates $p < 0.05$. The Steel-Dwass multiple comparison test was used as a statistical test. The best scores (excluding GOLD) are in bold.

	All		High		Low	
	Natural	Character	Natural	Character	Natural	Character
(a) AIML	2.71	2.44	2.74	2.42	2.49	2.63
(b) LUCENE	2.98^{aa}	3.13^{aa}	3.05^{aa}	3.13^{aa}	2.48	3.11
(c) PROP_WO_EXDB	3.04^{aa}	3.15^{aa}	3.09^{aa}	3.14^{aa}	2.62	3.19^{a}
(d) PROP	$\mathbf{3.23}^{aabbc}$	$\mathbf{3.24}^{aa}$	$\mathbf{3.28}^{aabb}$	$\mathbf{3.23}^{aa}$	**2.78**	$\mathbf{3.27}^{aa}$
(e) GOLD	$3.61^{aabbccdd}$	$3.74^{aabbccdd}$	$3.68^{aabbccdd}$	$3.75^{aabbccdd}$	3.11^{aabb}	3.65^{aab}

Table 5: Results for Ayase. See caption of Table 4 for notations in table.

Rule-based baseline (AIML) The typical approach to implement a chatbot is by using rules. We used the rules written in AIML created by Higashinaka et al (2015). There are roughly 300K rules. In Japanese, sentence-end expressions are key factors to exhibit personality. Therefore, following the method by Miyazaki et al. (2016), we created sentence-end conversion rules so that the output of this method would have the sentence-end expressions that match the characters in question.

Retrieval-based method (LUCENE) The retrieval-based method described in Section 3.1.

Proposed method 1 (PROP_WO_EXDB) The proposed method described in Section 3.2. This method does not use the extended question-answer pairs from Twitter. The weights $w_1 \ldots w_6$ are all set to 1.0. We used 10 for N for document retrieval.

Proposed method 2 (PROP) The proposed method with extended question-answer pairs from Twitter, as described in Section 3.4. We retrieved 10 candidates from collected question-answer pairs and 10 from extended ones. The weights $w_1 \ldots w_6$ are all set to 1.0.

Upper bound (GOLD) The gold responses by the online users to the test questions. When multiple answers are given to a question, one is randomly selected.

4.5 Results

Tables 4 and 5 list the results for Murai and Ayase, respectively. The topmost row indicates the level of knowledge about the characters. 'All' indicates the results of all participants, 'High' those who self-declared as being very knowledgeable, and 'Low' those who self-declared otherwise. We had 26 High and 6 Low participants for Murai, and 23 High and 3 Low participants for Ayase.

The tendencies were the same for the two characters, although the scores for Ayase were generally lower than those of Murai. AIML performed the worst followed by LUCENE. It was surprising that AIML's score was low; this is probably because of the peculiarities of the input questions for the characters. PROP_WO_EXDB and PROP performed better than AIML and LUCENE with statistical significance in many cases. GOLD was always the best-performing method. PROP was significantly better than PROP_WO_EXDB for naturalness but not for character-ness.

These results indicate that simple text-based retrieval is not sufficient, and we need more elaborate methods. The effectiveness of the extended question-answer pairs seems to be limited. It can be useful to make the interaction seem natural, but this does not necessarily improve character-ness, although we believe that having the ability to converse naturally is a requirement for chatbots.

When we focus on the results as they relate to the knowledge levels, we see large differences between High and Low. The High participants are likely to differentiate the answers more than Low

Question	Method	Answer	Nat.	Char.
What did you eat for lunch today?	AIML	I had a chicken breast rice bowl. It was good. What did you eat?	4.50	3.46
	LUCENE	Anything will do when you are playing video games.	1.88	2.08
	PROP_WO_EXDB	Ramen!	4.81	4.77
	PROP	Ramen.	4.73	4.31
	GOLD	I had a ginger grilled meal at Hidakaya.	4.92	4.15

Table 6: Examples of answers for Murai by each method. Questions and answers were originally in Japanese. English translations are by the authors.

Question	Method	Answer	Nat.	Char.
You are so cute.	AIML	Cute is the best.	2.69	2.15
	LUCENE	Shut up or I will sew your mouth.	3.04	4.12
	PROP_WO_EXDB	Thank you! It's a little embarrassing.	4.69	4.08
	PROP	Thank you!	4.62	4.15
	GOLD	You say silly things all the time. I may take it seriously.	4.08	4.27

Table 7: Examples of answers for Ayase by each method.

participants. For example, for Murai, there were only few cases in which there was statistical significance between the proposed methods when the knowledge level was low. The tendency was the same for Ayase. This highlights the difficulty in evaluating for characters.

Tables 6 and 7 show examples of answers for Murai and Ayase, respectively. Overall, since the proposed methods achieved character-ness scores well over 3 (which is the middle point in the scale), we conclude that we can create chatbots with consistent personalities by means of role play-based question-answering.

5 Related Work

Although there have not been any studies involving role play-based question-answering for data collection, there is a large body of research for creating chatbots that show consistent personalities.

There have been several studies on characters by generating or rewriting utterances reflecting the underlying personality traits (Mairesse and Walker, 2007; Sugiyama et al., 2014; Miyazaki et al., 2016). In addition, there has been extensive research on extending neural conversational models to reflect personal profiles (Li et al., 2016). Although such neural network-based methods show promising results, they still suffer from sparsity of data and non-informative utterances (Li et al., 2015). This paper proposed increasing the source data for character building; the data can be useful for neural models.

6 Summary and future work

Our goal for this study was to verify the effectiveness of role play-based question-answering for creating chatbots. Focusing on two famous characters in Japan, we successfully collected a large volume of question-answer pairs for two characters by using real users. We then created chatbots using the question-answer pairs. Subjective evaluation showed that although a simple text-retrieval based method does not work well, our proposed method that uses translation models as well as question-type matching and center-word extraction works well, showing reasonable scores in terms of naturalness and character-ness.

For future work, we need to consider approaches to improve the quality of the proposed method. For example, we are currently using equal weights for scoring. We believe that they can be optimized using training data. We also want to incorporate other pieces of information that may contribute to the ranking of answers, such as sentence embeddings (Kiros et al., 2015), discourse relations (Lin et al., 2009; Otsuka et al., 2017), and external knowledge about the characters. Although we used two very different characters in this paper, we want to use additional types of characters as targets for role play-based question-answering. We also want to incorporate the chatbots into the Web sites so that the users can feel they are training up the characters.

Acknowledgments

We thank the developers of DWANGO Co., Ltd. for creating the role play-based question-answering Web sites. We also thank the subscribers of the Max Murai and Tukasa Fushimi channels on NICONICO Douga for their cooperation. We thank the members of the Service Innovation Department at NTT DOCOMO, especially Yuiko Tsunomori, for helpful discussions and suggestions.

References

Reina Akama, Kazuaki Inada, Naoya Inoue, Sosuke Kobayashi, and Kentaro Inui. 2017. Generating stylistically consistent dialog responses with transfer learning. In *Proc. IJCNLP*, volume 2, pages 408–412.

Amanda Cercas Curry and Verena Rieser. 2016. A subjective evaluation of chatbot engines. In *Proc. WOCHAT*.

Carla Gordon, Jessica Tin, Jeremy Brown, Elisabeth Fritzsch, and Shirley Gabber. 2016. Wochat chatbot user experience summary. In *Proc. WOCHAT*.

Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. 2013a. Using role play for collecting question-answer pairs for dialogue agents. In *Proc. INTERSPEECH*, pages 1097–1100.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pages 928–939.

Ryuichiro Higashinaka, Toyomi Meguro, Hiroaki Sugiyama, Toshiro Makino, and Yoshihiro Matsuo. 2015. On the difficulty of improving hand-crafted rules in chat-oriented dialogue systems. In *Proc. APSIPA*, pages 1014–1018.

Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, and Nozomi Kobayashi. 2013b. Question answering technology for pinpointing answers to a wide range of questions. *NTT Technical Review*, 11(7).

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proc. NIPS*, pages 3294–3302.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2009. Building effective question answering characters. In *Proc. SIGDIAL*, pages 18–27.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proc. EMNLP*, pages 343–351.

François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proc. ACL*, pages 496–503.

Morry Van Ments. 1999. *The Effective Use of Role Play: Practical Techniques for Improving Learning*. Kogan Page Publishers.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pages 3111–3119.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 2016. Towards an entertaining natural language generation system: Linguistic peculiarities of Japanese fictional characters. In *Proc. SIGDIAL*, pages 319–328.

Atsushi Otsuka, Toru Hirano, Chiaki Miyazaki, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. 2017. Utterance selection using discourse relation filter for chat-oriented dialogue systems. In *Dialogues with Social Robots*, pages 355–365. Springer.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proc. LREC*.

Hiroaki Sugiyama, Toyomi Meguro, and Ryuichiro Higashinaka. 2017. Evaluation of question-answering system about conversational agent's personality. In *Dialogues with Social Robots*, pages 183–194. Springer.

Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. 2014. Large-scale collection and analysis of personal question-answer pairs for conversational agents. In *Proc. IVA*, pages 420–433.

Shota Takeuchi, Tobias Cincarek, Hiromichi Kawanami, Hiroshi Saruwatari, and Kiyohiro Shikano. 2007. Construction and optimization of a question and answer database for a real-environment speech-oriented guidance system. In *Proc. Oriental COCOSDA*, pages 149–154.

David Traum, Kallirroi Georgila, Ron Artstein, and Anton Leuski. 2015. Evaluating spoken dialogue processing for time-offset interaction. In *Proc. SIGDIAL*, pages 199–208.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Steve Walker, Stephen E Robertson, Mohand Boughanem, Gareth JF Jones, and Karen Sparck Jones. 1997. Okapi at TREC-6 automatic ad hoc, VLC, routing, filtering and QSDR. In *Proc. TREC*, pages 125–136.

Richard S Wallace. 2009. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer.