Response generation to out-of-database questions for example-based dialogue systems

Sota Isonishi, Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara

Abstract Example-based dialogue systems are often used in practice because of their robustness and simple architecture. However, when these systems are given out-of-database questions that are not registered in the question-response database, they have to respond with a fixed backup response, which can make users disengaged in the dialogue. In this study, we address response generation for out-of-database questions to make users perceive that the system understands the question itself. We define question types observed in the speed-dating scenario which is based on open-domain dialogue. Then we define possible response frames for each question type. We propose a sequence-to-sequence model that directly generates an appropriate response frame from an input question sentence in an end-to-end manner. The proposed model also explicitly integrates a question type classification to take into account the question type of the out-of-database question. Experimental results show that integrating the question type classification improved the response generation, and could exactly match 69.2% of response frames provided by human annotators.

1 Introduction

While various kinds of dialogue systems have been proposed and developed, example-based dialogue systems are still widely used as a simple, practical approach. Example-based dialogue systems use the database of pairs of an expected user question and the corresponding system response. When a question from the user is provided, an example-based dialogue system retrieves the nearest question (or the most related response) from a database using methods such as keyword matching or vector space modeling, and the corresponding response will be generated

Graduate School of Informatics, Kyoto University, Japan
e-mail: [isonishi][inoue][lala][takanashi][kawahara]@sap.ist.i.kyoto-u.ac.jp

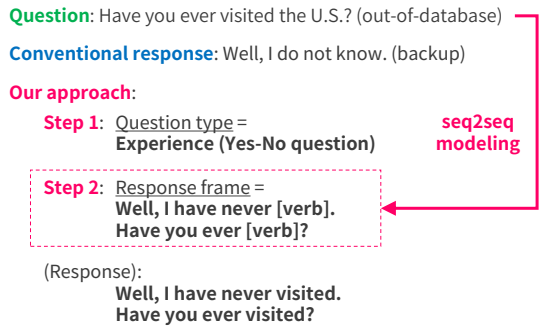

Fig. 1 Overview of response generation to out-of-database questions

as the system response. Recently, the matching part has massively adopted neural networks [15, 7, 16, 18, 19, 17, 20]. Those works have made the example-based dialogue systems more robust and also made it possible to increase the scale of the database.

The obvious limitation of example-based dialogue systems is that it is impossible to answer out-of-database questions - questions that are not registered in the database. In this case, the system has to generate the above-mentioned safe responses, which might degrade the naturalness of dialogue and decrease user engagement. Although it is important for example-based dialogue systems to register as many expected questions as possible, it is not practical to cover all possible questions, especially in practical dialogue systems that deal with a wide range of domains. Therefore, we need an effective approach to generate proper responses for out-of-database questions in order to retain the naturalness of dialogue and user engagement.

We propose response generation to out-of-database questions to make users perceive that the system is not able to answer but is able to understand the question itself. Our approach is summarized in Fig. 1. First, we classify the question type of the input question (Step 1). For example, the question type of "Have you been to the U.S.?" is "experience". Question type classification has been studied as a key function in conventional question answering systems [12, 10, 11, 9]. In this study, we design a set of question types in the domain of speed-dating which is a dialogue of first encounters. Since it is open-domain dialogue, the proposed question type has the potential to be extended to other kinds of dialogue. Next, we choose from pre-defined response frames according to the classified question type to generate a response sentence (Step 2). Since the question type classification is error-prone and there are multiple possible response frames, we propose an approach that generates a proper response. In this study, we apply a neural network-based sequence-to-sequence (seq2seq) model to directly generate a response frame from an input question sentence in the manner of end-to-end modeling. To take into account question type classification, we conduct pre-training of question type classification independently and integrate the pre-trained model to the seq2seq model. This study contributes to realizing example-based dialogue systems that are robust against out-of-database questions and continue dialogue without breakdown.

Fig. 2 Snapshot from speed-dating dialogue corpus (left: subject and ERICA, right: operator)

2 Human-robot speed-dating dialogue corpus

We use a human-robot dialogue corpus where a human subject talked with an android robot ERICA [2, 3]. From this corpus, we chose one of the dialogue tasks, speed-dating which is an open-domain dialogue. In this task, participants meet each other and exchange their profile information to build the relationship between them. ERICA plays the role of a conversation practice partner for a male subject.

ERICA was operated by another human subject, called an operator. Fig. 2 shows a snapshot of the dialogue recording. The operator's voice was directly played by a speaker placed on ERICA. The operator manually controlled non-verbal behaviors of ERICA such as eye gaze, head nodding, and hand gesture. With this setting, 31 dialogue sessions were recorded where each session lasted about 10 minutes. Whereas the subject was a different person in each session, the operator was one of four trained actresses. In advance, both operator and subject were given a list of dialogue topics which are generally talked about in first encounter dialogues.

For the current study, we extracted question sentences from this corpus. At first, we annotated dialogue act labels based on the standard definition [1] to recognize questions uttered by both ERICA and the subjects. To simplify the problem in this study, we use only questions that do not require the dialogue context to understand the meaning of the question. We call these questions base questions. On the other hand, questions that require additional dialogue context such as "Where was it?" were excluded from the current scope. The number of the base questions in the corpus was 370.

3 Question type

We now define question types and report a manual annotation for the speed-dating corpus including augmented data.

Table 1 Definition and example of question types

Question type	Definition
In-database	Questions about personal information of the system which can be searched in our database and frequently asked in first-encountering dialogue e.g. "What is your hometown?", "What is your hobby?"
Habit	Questions about actions regularly done by the system e.g. "What do you do in your day off?"
Preference	Questions about preferences except those directly asking about hobbies e.g. "Do you like traveling?"
Experience	Questions about past actions e.g. "Have you ever been to the U.S.?"
Desire	Questions about future plans and hopes e.g. "How will you spend the next summer vacation?"
Subjective thoughts	Questions about subjective thoughts on something e.g. "Do you feel worthwhile to work?"
Personal information	Questions about personal information of the system which is not contained in our database e.g. "Do you have any pets?"
General knowledge	Questions about general knowledge, not related to the system e.g. "Do you know the Uyuni salt lake?"

3.1 Definition of question type

We analyzed the base questions observed in the corpus and defined question types. The question types consist of *in-database*, *habit*, *preference*, *experience*, *desire*, *subjective thoughts*, *personal information*, and *general knowledge*. The definition and example of the question types are summarized in Table 1. Since the speed-dating corpus covers many topics for exchanging their profiles in open-domain, this set of question types will be applied to other kinds of social dialogue. Currently we have a database of frequently used question-answer pairs for ERICA, based largely on previous user interactions. We assume that *in-database* question types can be searched for within this database. Therefore, *in-database* questions are out of scope in the current study

Furthermore, we took into account the form of questions because the corresponding response frames are different, which is explained in the next section. Based on the dialogue act labels, we classified the base questions into two forms: propositional questions (Yes-No) and set questions (5W1H). In this study, we consider the combination of question form (Yes-No or 5W1H) and the question type. We term this combination a *question type condition*.

Table 2 Distribution of question types

Question type	Question form		Total
	Yes-No	5W1H	
Habit	115	120	235
Preference	100	122	222
Experience	97	99	196
Desire	58	80	138
Subjective thoughts	67	54	121
Personal info.	60	57	117
General knowledge	81	58	139
Total	578	590	1,168

3.2 Annotation of question type

We manually annotated the question type condition for the base questions observed in the speed-dating corpus. Since the number of question samples in the corpus is not enough for machine learning, we conducted data augmentation. We asked third-party augmenters to create base question sentences, giving the combination of a dialogue topic and a question type condition. Note that *in-database* questions were not The given dialogue topics are based on those observed in the speed-dating corpus. Each augmenter was given one of 14 question type conditions, excepting *in-database*. For example, when an augmenter was given a topic of *travel* and a question form of *Yes-No*, they created a question sentence like *Do you like traveling since you were young?*. We recruited four people as the augmenters and obtained 923 additional base question sentences.

Table 2 reports the distribution of the questions types. The total number of questions was 1,168. Note that *in-database* question are not included here. In the next section, we also annotate appropriate response frames for each question sentence. As a result, some question sentences did not correspond to any response frames in the viewpoint of context validity, so in later experiments we use 1,109 question sentences that were associated with adequate response frames. Although we regard all the 1,109 questions as out-of-database questions in the current study, this scope depends on the database of example-based dialogue systems.

3.3 Question type classification

We conducted a preliminary experiment on question type classification. The input is a bag-of-words feature of function words, pronouns, adverbial nouns and adjective verbs. Note that we used only words that appeared in more than 1% of the training data. Furthermore, we also used a flag indicating whether the verb in the past tense. The output is a question type condition (14 types). Therefore, the current task is 14-class classification and we implemented this using a logistic regression model.

Table 3 Result of question type classification

Question type	Question form					
	Yes-No			5W1H		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Habit	0.938 (106 / 113)	0.964 (106 / 110)	0.872	0.763 (87 / 114)	0.757 (87 / 115)	0.760
Preference	0.957 (88 / 92)	0.880 (88 / 100)	0.917	0.787 (107 / 136)	0.856 (107 / 125)	0.820
Experience	0.750 (75 / 100)	0.833 (75 / 90)	0.789	0.736 (67 / 91)	0.744 (67 / 90)	0.740
Desire	0.660 (33 / 50)	0.600 (33 / 55)	0.629	0.652 (45 / 69)	0.600 (45 / 75)	0.625
Subjective thoughts	0.754 (52 / 69)	0.800 (52 / 65)	0.776	0.789 (30 / 38)	0.600 (30 / 50)	0.682
Personal information	0.518 (29 / 56)	0.580 (29 / 50)	0.547	0.472 (17 / 36)	0.378 (17 / 45)	0.420
General knowledge	0.782 (68 / 87)	0.850 (68 / 80)	0.814	0.641 (25 / 39)	0.417 (25 / 60)	0.505

We evaluated the above model with 5-fold cross validation of the 1,109 question sentences. Table 3 reports the classification result. The overall accuracy of 0.746 was much higher than chance level (0.112) which classifies to the majority class (*preference*, 5W1H). We observed especially high scores on well-observed question types such as *habit*, *preference*, and *experience*. This result suggests the practicality of the question type classification in the current dialogue task.

4 Response frame

For each question type, we designed several response frames that make users perceive that the system is not able to answer but is able to understand the question itself. Since there are several possible response frames for each question type, we then manually annotated appropriate frames for each question sentence.

4.1 Design of response frames

We manually designed several response frames for each question type condition (the combination of a question form and a question type). For example, if a question sentence is " Do you often watch movies? ", the question form is *Yes-No* and the question type is *habit*. One of the response frames is "Well, I do not [verb] [focus]. Do you often [verb] [focus]?". The full set of response frames is summarized in

Table 4 Response frame for Yes-No question form (✓represents that the corresponding item, e.g. a focus word, must be found in an input question.)

Question type	Focus	Verb	Response frame
Habit	✓	✓	Well, I do not [verb]. Do you often [verb] [focus]?
	✓	-	Well, I do not do that. Do you often do [focus]?
	-	✓	Well, I do not [verb]. Do you often [verb]?
	-	-	Well, I do not do that. Do you often do that?
Preference	✓	-	Well, not really. Do you like [focus]?
	-	-	Well, not really. Do you like that?
Experience	-	✓	Well, I have never [verb]. Have you ever [verb]?
	-	✓	Well, I did not [verb]. Did you [verb]?
	-	-	Well, I have never done that. Have you ever done that?
	-	-	Well, I did not do that. Did you do that?
Desire	-	✓	Well, I think nothing in particular. Will you [verb]?
	-	-	Well, I think nothing in particular. How about you?
Subjective thoughts	-	-	Well, not really. How about you?
Personal information	-	-	Well, I am not sure about that. How about you?
	-	-	Well, nothing special. How about you?
	-	-	Well, not really. How about you?
General knowledge	-	-	Well, I do not know that. Do you know that well?

Table 4 and Table 5. Note that the response frames are designed in the Japanese language.

Each response frame consists of two parts: *reaction* and *question*. *Reaction* is the first part of the response frame and is used to return a denial answer to a user question. This denial answer prevents any elaboration on the out-of-database question that cannot be handled by the current example-based dialogue system. *Question* is the second part of the response frame, and is used to ask a question back to the user. Therefore, the system can take the dialogue initiative. In the above example, *reaction* corresponds to "Well, I do not [verb] [focus]." and *question* corresponds to "Do you often [verb] [focus]?". When the dialogue system generates a response sentence based on this response frame, it is expected to make the user perceive that the system understood the question itself, and then also make it engaged in the dialogue. In this work, the proposed method generates response frames including slots without filling specific words in slots. The method to extract focus and verb words is left in future work, though it is possible to use tools from previous studies on morphological analysis [4] and focus word detection [5].

In each question type condition, we needed to create several response frames to cover the expected variation of the input question sentences. For example, on the question type condition of *preference* and *5W1H*, we created two kinds of response frames. The appropriate response frame depends on the input question sentence. If the input question is "What kind of food do you like?", the response frame should be "Well, nothing in particular. What do you like?". On the other hand, if the question is "Which football players do you like?", the response frame should b "Well, nobody in particular. Who do you like?". Furthermore, we also created several response

Table 5 Response frame for 5W1H question form (✓represents that the corresponding item, e.g. a verb, must be found in an input question. WH represents an interrogative.)

Question type	Verb	WH	Response
Habit	✓	✓	Well, there it nothing [WH]. [WH] do you [verb]?
	✓	-	Well, I do not [verb] in particular. What do you [verb]?
	-	✓	Well, there is nothing [WH]. [WH] do you do?
	-	-	Well, I do not do in particular. What do you do?
Preference	-	-	Well, nothing in particular. What do you like?
	-	-	Well, nobody in particular. Who do you like?
Experience	✓	-	Well, I did not [verb] in particular. How about you?
	✓	-	Well, I did not [verb]. Did you [verb] anything?
	-	-	Well, I did nothing in particular. How about you?
	-	-	Well, I did not do that. Did you do anything?
Desire	✓	-	Well, I think nothing in particular. Is there any places you want to [verb]?
	✓	-	Well, I think nothing in particular. Is there anything you want to [verb]?
	-	-	Well, I think nothing in particular. Do you have any plan?
	-	-	Well, I think nothing in particular. Is there anything?
Subjective thoughts	-	-	Well, I do not think anything in particular. How about you?
Personal information	-	-	Well, I am not sure about that. How about you?
	-	-	Well, there are various. How about you?
	-	-	Well, nothing special. How about you?
General knowledge	-	-	Well, I do not know. Do you know that well?

frames depending on the presence of slot words. For example, on the question type condition of *preference* and *Yes-No*, we created two kinds of response frames depending on the presence of a focus word inside the question. If there is a focus word inside the question such as “Do you like football?”, the response frame should be “Well, not really. Do you like [focus = football]?” If a focus word is ambiguous, the response frame can be “Well, not really. Do you like that?”. From the above, there is no one-to-one relationship between the question type and the response frame.

4.2 Annotation of corresponding response frames

We annotated appropriate frames for each question sentence. As each question sentence was already associated with a question type condition, we checked if each response frame candidate was appropriate in the viewpoint of the dialogue context and the presence of slot words inside the question sentence. Note that the number of appropriate frames was not restricted to one, which means that in some cases there are several frames for an input question sentence. For this question type condition, there are four response frames, but all frames contain a word “often” inside the response frame. This word is inconsistent with the meaning of the question sentence,

so this question was annotated as there was no appropriate response frame. As a result, 1,109 of 1,168 question sentences were associated with more than one response frame. In this study, we do not use the 59 samples that were not associated with any frames. We use the pairs of the input question sentences and associated appropriate response frames in order to train a sequence-to-sequence model in the next section.

5 Response frame generation

To generate the appropriate response frame, we need to classify the question type condition and then find the appropriate response frame from several candidates. Since the question type classification is error-prone and there is ambiguity in the mapping from the question type condition to the appropriate frame, we propose sequence-to-sequence (seq2seq) modeling to generate the response frame directly from the input out-of-database question sentence. First, we introduce a simple seq2seq model. Then, we integrate the question type classification model in order to take into account which question type condition the input question sentence is.

5.1 Simple seq2seq model

The first model is a simple seq2seq model implemented by recurrent neural networks [14]. We use gated recurrent units (GRU) in the current study. The input is a word sequence of a question sentence and the output is a word sequence of a response frame annotated in the previous section. The input word is embedded in a distributed representation (word2vec) by using a continuous bag-of-word (CBOW) model [6]. This word embedding is trained independently from the training of the seq2seq model by using the same training data. Note that we gave a random vector of zero-mean and unit-variance for unknown words that did not appear in the training data. The output word sequence is searched using a greedy method. An input sequence sometimes corresponds to several output sequences when several response frames were annotated as adequate for the same input question sentence. In this case, we regard that there are several data pairs with the same input sequence and different output sequences, and we use them individually.

The seq2seq model has the potential to be able to handle two tasks, the question type classification and the response frame selection, with a simple architecture in an end-to-end manner. Another advantage of using the seq2seq model for the current task is the extendability of the model. Even if we use more variations of output response frames, the seq2seq model can easily handle the increase of the variation of output sequences. On the other hand, if we use another model that selects the kind of output response frame instead of generating a response frame itself, the output dimension would increase with the number of response frames, which is inefficient. In future work, we also plan to directly generate a response sentence, automatically

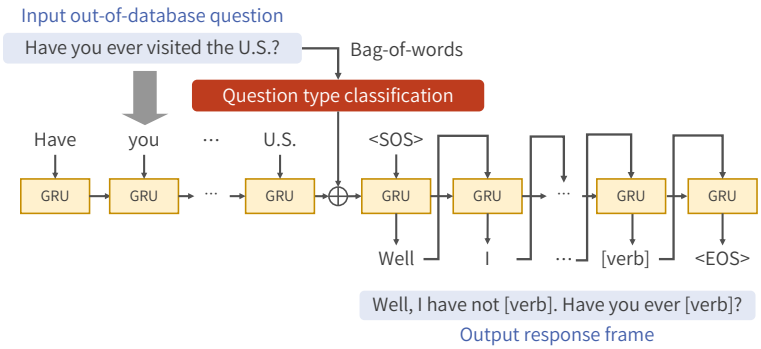

Fig. 3 Seq2seq model generating a response frame by taking into account question type classification

filling in the slot words. In this case, one possible and sophisticated way is to extend the seq2seq model to something like a pointer network which directly generates a response sentence by referring to some words in the input sequence [13, 8].

5.2 Integration with question type classification

We extend the simple seq2seq model. In our preliminary experiment in Section 3.3, we confirmed that we can classify the question type with a reasonable accuracy. The information about question type is useful for the seq2seq model to generate an adequate response frame. Accordingly, we propose to integrate the question type classification into the seq2seq model. Figure 3 illustrates the architecture of the proposed model. We concatenate the output of the logistic regression of the question type classification (14 dimensions) and the output of the encoder of the seq2seq and then feed the concatenated vector to the decoder. In training, we pre-train the logistic regression of question type classification and fine-tune the entire network. This integrated network is expected to generate a response frame which considers the question type.

6 Evaluation

We evaluated response generation to out-of-database questions with the seq2seq models.

Table 6 Evaluation of response generation

Method	Perfect matching ratio
Baseline	0.635
Simple seq2seq	0.672
Integrated seq2seq	0.692

6.1 Setup

We conducted 5-fold cross validation of the 1,109 question sentences. Since each input question sentence can correspond to more than one output response frame, the number of training data pairs was 1,016. The evaluation metric was an perfect matching between a generated response frame and a reference response frame. We regarded that it was correct if the generated response frame matched one of the corresponding reference response frames.

We also implemented a baseline method that is based on the question type classification method described in section 3.3. This model first classifies the question type condition and consequently select a response frame that is the most frequent one in the training dataset within the classified question type condition. We compared the two seq2seq models with this baseline model.

The parameters of the seq2seq models are as follows. We use one layer of GRUs for the encoder and decoder of the seq2seq models. The number of hidden units of the GRUs was 128. Note that when we integrated the question type classification, the number of hidden units of the decoder increased to 142 (128+14). The dropout ratio was 20%. We regarded one sequence as one minibatch. The number of training epochs was 100.

6.2 Result

Table 6 reports the ratios of perfect matching between the generate response frame and the reference response frame. Compared with the baseline model, the seq2seq models showed higher scores, which demonstrates the effect of this end-to-end modeling. The integrated seq2seq model increased the matching ratio to 0.692 from 0.672. This improvement shows that it is effective to explicitly take into account the kind of question type condition in the current response generation task.

6.3 Generated examples

We report several generated examples by the seq2seq models. Examples have been translated from Japanese, so the grammar of these may not be necessarily correct

in English. Correct examples of both seq2seq models are shown below. Note that **Q** and **R** are an input question sentence and a generated response frame, respectively.

Q1 Do you go to music concerts? (Habit, Yes-No)
R1 Well, I do not [verb]. Do you often [verb] [focus]?

Q2 What do you do on your days off? (Habit, 5W1H)
R2 Well, there is nothing [interrogative] I do in particular. [interrogative] do you [verb]?

Q3 Have you ever kept any pets? (Experience, 5W1H)
R3 Well, I did not [verb] in particular. How about you?

Some examples where the simple seq2seq model failed but the integrated one succeeded are follows. Note that **Si** is the generated response frames by the simple seq2seq and **In** is one generated by the integrated seq2seq model, respectively.

Q4 Do you know a music instrument called a Cajon? (General knowledge, Yes-No)
Si4 Well, there it nothing [interrogative]. [interrogative] do you [verb]? (Habit, 5W1H)
In4 Well, I do not know. Do you know that well?

Q5 What do you do when you meet your family? (Habit, 5W1H)
Si5 Well, I think nothing in particular. How about you? (Desire, 5W1H)
In5 Well, there is nothing [interrogative]. [interrogative] do you [verb]?

In the above examples, the generated response frame by the simple seq2seq frame came from an incorrect question type condition. Since the integrated seq2seq model explicitly takes into account the kind of question type condition, these cases were enhanced.

6.4 Human evaluation

We conducted a human evaluation to confirm the appropriateness of the generated response frames. We recruited three human evaluators. From the response frames generated by the integrated seq2seq model, we randomly selected 100 samples keeping the balance of correct and wrong samples. We showed each pair of the input question sentence and the generated response frame and then asked each evaluator to judge if he/she feels that the system understands the question itself. Note that we also asked the evaluators to interpolate adequate slot words by themselves as much as possible. The evaluator had to select one of three choices: *accept* (the system understands the question), *not accept* (not understand), or *neither*. We used the majority voting that regarded the sample as *accepted by human* if more than two persons selected *agree* for the sample.

Table 7 reports the result of the human evaluation. In total, 73 of 100 samples were accepted by the evaluators, slightly higher than the perfect matching ratio. In the correct samples, 57 of 69 samples were accepted by human. The remaining 12

Table 7 Result of human evaluation

		Human evaluation		Total
		Accept	No	
Generation	Correct	57	12	69
	Wrong	16	15	31
Total		73	27	100

samples were not accepted even though the response frame was correctly generated. This suggests that the set of response frames needs to be revised in future work. Within the wrong samples, surprisingly, 16 of 31 samples were accepted. When we analyzed these samples they seemed to be grammatically incorrect but semantically acceptable and meaningful. These types of response can be accepted and used in conversations such as speed-dating dialogue where engagement is more important than the grammatical correctness.

7 Conclusion

We have addressed response generation to out-of-database question for example-based dialogue systems to make users perceive that the system understands the question itself. We defined question types such as *habit* and *preference* based on observation of speed-dating first-encountering dialogue which is open domain. We then designed response frames for each combination (called question type condition) of the question type and the question form (Yes-No or 5W1H). The sequence-to-sequence neural network model was used to directly generate the adequate response frame from the input question sentence. Meanwhile, we confirmed that we could classify the question type condition with an accuracy of 0.746. Therefore, we integrated the question type classification (logistic regression) into the seq2seq model to explicitly consider the question type condition. This model achieved an exact sequence matching ratio of 0.692, improving the basic seq2seq model. Finally, we confirmed that human evaluators accepted 73 of 100 generated samples in the viewpoint of whether the system understands the question itself.

Our future work is as follows. In the current study, our model generates response frames that contains slots for entries such as focus and verb words. To fill in the slots, we will first use conventional tools such as morphological analysis to extract the slot values. As a more sophisticated approach, we will investigate how to extract the slot values by using an integrated neural network such as pointer networks [13, 8]. We will also conduct a subjective evaluation for the generated response sentences including slot values. Finally, we will integrate the response generate to out-of-database questions into current example-based dialogue systems to apply them in practice.

Acknowledgments

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant number JPMJER1401) and Grant-in-Aid for Scientific Research on Innovative Areas "Communicative intelligent systems towards a human-machine symbiotic society" (Grant number JP19H05691).

References

1. Bunt, H., Alexandersson, J., Carletta, J., Choe, J.W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., Traum, D.: Towards an ISO standard for dialogue act annotation. In: LREC, pp. 2548–2555 (2010)
2. Inoue, K., Milhorat, P., Lala, D., Zhao, T., Kawahara, T.: Talking with ERICA, an autonomous android. In: SIGDIAL, pp. 212–215 (2016)
3. Kawahara, T.: Spoken dialogue system for a human-like conversational robot ERICA. In: IWSDS (2018)
4. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to japanese morphological analysis. In: EMNLP, pp. 230–237 (2004)
5. Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., Kawahara, T.: Attentive listening system with backchanneling, response generation and flexible turn-taking. In: SIGDIAL, pp. 127–136 (2017)
6. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: ICML, pp. 1188–1196 (2014)
7. Lowe, R., Pow, N., Serban, I., Pineau, J.: The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In: SIGDIAL, pp. 285–294 (2015)
8. Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. In: ICLR (2017)
9. Mitamura, T., Lin, F., Shima, H., Wang, M., Ko, J., Betteridge, J., Bilotti, M.W., Schlaikjer, A.H., Nyberg, E.: Javelin iii: Cross-lingual question answering from japanese and chinese documents. In: NTCIR (2007)
10. Mizuno, J., Akiba, T., Fujii, A., Itou, K.: Non-factoid question answering experiments at ntcir-6: Towards answer type detection for realworld questions. In: NTCIR (2007)
11. Shima, H., Mitamura, T.: Javelin iii: Answering non-factoid questions in japanese. In: NTCIR (2007)
12. Tamura, A., Takamura, H., Okumura, M.: Classification of multiple-sentence questions. In: IJCNLP, pp. 426–437 (2005)
13. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: NIPS, pp. 2692–2700 (2015)
14. Vinyals, O., Le, Q.: A neural conversational model. In: ICML Deep Learning Workshop (2015)
15. Wang, H., Lu, Z., Li, H., Chen, E.: A dataset for research on short-text conversations. In: EMNLP, pp. 935–945 (2013)
16. Wang, M., Lu, Z., Li, H., Liu, Q.: Syntax-based deep matching of short texts. In: IJCAI, pp. 1354–1451 (2015)
17. Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z.: Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In: ACL, pp. 496–505 (2017)
18. Yan, R., Song, Y., Wu, H.: Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In: SIGIR, pp. 55–64 (2016)
19. Zhou, X., Dong, D., Wu, H., Zhao, S., Yu, D., Tian, H., Liu, X., Yan, R.: Multi-view response selection for human-computer conversation. In: EMNLP, pp. 372–381 (2016)
20. Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W.X., Yu, D., Wu, H.: Multi-turn response selection for chatbots with deep attention matching network. In: ACL, pp. 1118–1127 (2018)