A conversational dialogue manager for the humanoid robot ERICA

Pierrick Milhorat, Divesh Lala, Koji Inoue, Tianyu Zhao, Masanari Ishida, Katsuya Takanashi, Shizuka Nakamura and Tatsuya Kawahara

Abstract We present a dialog system for a conversational robot, Erica. Our goal is for Erica to engage in more human-like conversation, rather than being a simple question-answering robot. Our dialogue manager integrates question-answering with a statement response component which generates dialog by asking about focused words detected in the user's utterance, and a proactive initiator which generates dialog based on events detected by Erica. We evaluate the statement response component and find that it produces coherent responses to a majority of user utterances taken from a human-machine dialog corpus. An initial study with real users also shows that it reduces the number of fallback utterances by half. Our system is beneficial for producing mixed-initiative conversation.

1 Introduction

Androids are a form of humanoid robots which are intended to look, move and perceive the world like human beings. Human-machine interaction supported with androids has been studied for some years, with many works gauging user acceptance of tele-operated androids in public places such as outdoor festivals and shopping malls [3, 17]. Relatively few have the ability to hold a multimodal conversation autonomously, one of these being the android Nadine [26].

In this paper, we introduce a dialogue management system for Erica (ERato Intelligent Conversational Android). Erica is a Japanese-speaking android which converses with one or more human users. She is able to perceive the environment and users through microphones, video cameras, depth and motion sensors. The design objective is for Erica to maintain an autonomous prolonged conversation on a variety of topics in a human-like manner. An image of Erica is shown in Fig. 1.

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

Fig. 1 The android Erica is designed to be physically realistic. Motors within her face provide speaking motions in addition to unconscious behaviors such as breathing and blinking.

Erica's realistic physical appearance implies that her spoken dialogue system must have the ability to hold a conversation in a similarly human-like manner by displaying conversational aspects such as backchannels, turn-taking and fillers. We want Erica to have the ability to create mixed-initiative dialogues through a robust answer retrieval system, a dynamic statement-response generation and proactively initiating a conversation. This distinguishes Erica as a conversational partner as opposed to smartphone-embedded vocal assistants or text-based chatbot applications. Erica must consider many types of dialogue so it can take on a wide range of conversational roles.

Chatting systems, often called chatbots, conduct a conversation with their users. They may be based on rules [8, 4] or machine-learnt dialogue models [21, 25]. Conducting a conversation has a wide meaning for a dialogue system. Wide variations exist in the modalities employed, the knowledge sources available, the embodiment and the physical human-likeness. Erica is fully embodied, ultra realistic and may express emotions. Our aim is not for the dialogue system to have access to a vast amount of knowledge, but to be able to talk and answer questions about more personal topics. She should also demonstrate attentive listening abilities where she shows sustained interest in the discourse and attempts to increase user engagement.

Several virtual agent systems are tuned towards question-answering dialogues by using information retrieval techniques [22, 15]. However these techniques are not flexible enough to accept a wide variety of user utterances other than well-defined queries. They resort to a default failing answer when unable to provide a confident one. Moreover most information retrieval engines assume the inputs to be text-based or a near-perfect speech transcription. One additional drawback of such an omniscient system is the latency they introduce in the interaction when searching for a response. Our goal is to avoid this latency by providing appropriate feedback even if the user's dialogue is uncertain.

Towards this goal we introduce an attentive listener which convinces the interlocutor of interest in the dialogue so that they continue an interaction. To do this, a system typically produces backchannels and other feedback with the limited understanding it has of the discourse. Since a deep understanding of the context of the dialogue or the semantic of the user utterance is unnecessary, some automatic attentive listeners have been developed as open domain systems [10]. Others actually use

predefined scripts or sub-dialogues that are pooled together to iteratively build the ongoing interaction [1]. One advantage is that attentive listeners do not need to completely understand the user's dialogue to provide a suitable response. We present a novel approach based on capturing the focus word in the input utterance which is then used in an n-gram-based sentence construction.

Virtual agents combining good question-answering abilities with an attentive listeners are rare. SimSensei Kiosk [23] is an example of a sophisticated agent which integrates backchannels into a virtual interviewer. The virtual character is displayed on a screen and thus does not situate the interaction in the real world. Erica iteratively builds a short-term interaction path in order to demonstrate a mixed-initiative multimodal conversation. Her aims is to keep the user engaged in the dialogue by answering questions and showing her interest. We use a hierarchical approach to control the system's utterance generation. The top-level controller queries and decides on which component (question-answering, statement response, backchannel or proactive initiator) shall take the lead in the interaction.

This work presents the integration of these conversation-based components as the foundation of the dialogue management system for Erica. We introduce the general architecture in the next section. Within section 3, we describe the individual components of the system and then conduct a preliminary evaluation in Section 4. Note that Erica speaks Japanese, so translations are given in English where necessary.

2 Architecture

Erica's dialogue system combines various individual components. A top-level controller selects the appropriate component to use depending on the state of the dialogue. We cluster dialogue segments into four main classes as shown in Table 1 (examples are translated from Japanese). The controller estimates the current dialogue segment based on the short-term history of the conversation and then triggers the appropriate module to generate a response.

Fig. 2 shows a general architecture of Erica's dialogue system, focusing on the speech and event-based dialogue management which is the topic of this paper.

The system uses a Kinect sensor to reliably identify a speaker in the environment. Users' utterances are transcribed by the speech recognizer and aligned with a tracked user. An event detector continuously monitors variables of the space and sound environment to extract selected events such as periods of silence and user locations. An interaction process is divided in steps, triggered by the reception of an input which is either a transcribed utterance or a detected event, and ends with a decision/selection.

First, the received utterance is sent to the question-answering and statement response components which generate an associated confidence score. This score is based on factors such as the hypothesized dialogue act of the user utterance and the presence of keywords and focus phrases. The controller then selects the component's response with the highest confidence score. However if this score does not

Table 1 Classification of dialogue segments ([] = event/action, "" = utterance)

Class	Example	Component(s) involved
Question-answering	U: "What is your name?" S: "I'm Erica"	Question Answering Proactive initiator Backchanneling
Attentive listening	U: "I went to Osaka yesterday" S: "Hum." [nodding] S: "What did you do in Osaka?"	Statement response Backchanneling
Topic introduction	U/S: [4-second silence] S: "Hey, do you like dancing?"	Proactive initiator
Greetings/Farewell	U: [entering social space] S: "Hello, I am Erica, would you like to talk a bit?"	Proactive initiator

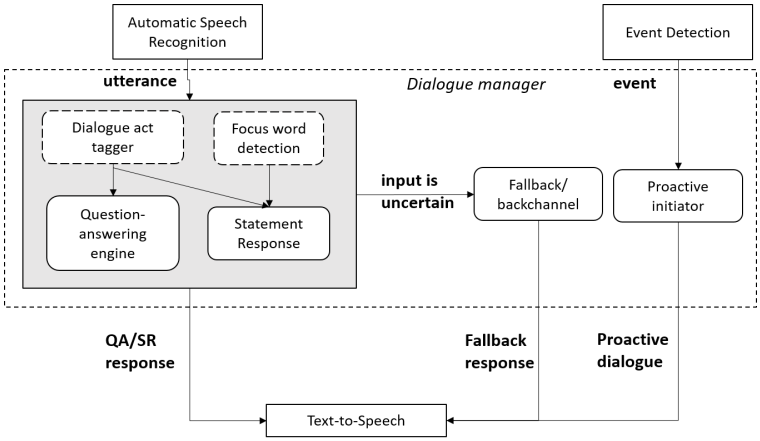

Fig. 2 Architecture of the dialogue system

meet the minimum threshold, the dialogue manager produces a backchannel fallback.

Both the question-answering and statement response components use dialogue act tagging to generate their confidence scores. We use a dialogue act tagger based on support vector machines to classify an utterance into a questions or non-question. Focus word detection is used by the statement response system and is described in more detail in Section 3.2.

Events such as silences and users entering the proximity of Erica are detected by and handled by the proactive initiator. Erica instigates dialogue which is not in response to any speech input from the user. This dialogue is generated based on rules and is described in more detail in Section 3.3.

3 Components

In this section, we describe individual components of the system and example responses they generate.

3.1 Question answering with fallback

Task-oriented spoken dialogue systems handle uncertain inputs with explicit or implicit confirmations [9]. There is a trade-off between the consequences of processing an erroneous utterance and the expected fluency of the system [24]. Question-answering engines such as smartphone assistants make no confirmations and let users decide whether they accept the returned results. As a conversational partner, Erica cannot use such explicit strategies as they interrupt the flow of the dialogue. We can consider chatting with uncertainty to be similar to conversing with non-native speakers, with misunderstandings being communicated and repaired jointly.

Erica's question-answering components enables her to implicitly handle errors and uncertainty. Since the system's goal is to generate conversational dialogues, an exact deep understanding of the user utterances is not necessary. Erica is able to generate implicit non-understanding prompts such as "e?" ("huh?" in English), backchannels and nodding. These signals are used when the system is unable to generate an answer with sufficiently high confidence.

The following conversation shows an instance of interaction segment between a user and Erica in which her responses are managed with only the question-answering and the backchannel components:

Table 2 Example of question-answering based interaction ([] = event/action, "" = utterance)

U: "Where does your name come from?"
S: "My name is spelled E R I C A. It stands for ERato Intelligent Conversational Android [...]." (Question Answering)
U: "That is a strange name for a robot, isn't it?"
S: [nodding] (Backchanneling)
U: "I am from Osaka. Do you know about Abeno Harukas?"
S: "Abeno Harukas is a tower in Osaka which [...]?" (Question Answering)

The question-answering manager bases its knowledge on a handcrafted database of adjacency pairs. The following measure is used to compare a set of n ranked speech recognition hypotheses $\{(u_1, cm_1), (u_2, cm_2)...(u_n, cm_n)\}$ and all first-pair parts fpp_{db} in the database:

$$m(u_i, cm_i, fpp_{db}) = \frac{1}{1 + e^{\alpha.ld(fpp_{db}, u_i) + \beta.(1 - cm_i) + \gamma}}$$

$ld(fpp_{db}, u_i)$ is the normalized Levenshtein distance between a database entry fpp_{db} and the hypothesis' utterance u_i. cm_i is the confidence measure of the speech recognizer mapped to the interval $[0;1]$ using the sigmoid function. α and β are weights given to the language understanding and speech recognition parts. γ is a bias that determines the overall degree of acceptance of the system. This approach is not highly sophisticated, but is not the main focus of this work. We found it sufficient for most user questions which were within the scope of the database of topics.

The algorithm searches for the most similar first-pair part given the incoming input. The entry for which the computed measure is the lowest is selected and the associated system response is generated. If the measure m does not exceed a threshold, the system resorts to a fallback response.

3.2 Statement response

In addition to answering questions from a user, Erica can also generate focus-based responses to statements. Statements are defined as utterances that do not explicitly request the system to provide information and are not responses to questions. For instance, "Today I will go shopping with my friends" or "I am happy about your wedding" are statements. Chatting is largely based on such exchanges of information, with varying degrees of intimacy depending on speaker familiarity.

Higashinaka et al. [11] proposed a method to automatically generate and rank responses to why-questions asked by users. Previous work also offered a similar learning method to help disambiguate the natural language understanding process using the larger dialogue context [12, 14] and to map from semantic concept to turn realization [13].

Our approach is based on knowledge of common statement responses in Japanese conversation [7]. This includes some repetition of the utterance of the previous speaker, but does not require full understanding of their utterance. As an example, consider the user utterance "Yesterday, I ate a pizza". Erica's objective is to engage the user and so may elaborate on the question ("What kind of pizza?") or partially repeat the utterance with a rising tone ("A pizza?"). The key is the knowledge that "pizza" is the most relevant word in the previous utterance. This has also been used in previous robot dialogue systems [19].

We define four cases for replying to a statement as shown in Fig. 3, with examples shown in Table 3. Focus phrases and predicates are underlined and question words are in boxes. Similar to question-answering, a fallback utterance is used when no suitable response can be found, indicated in the table as a formulaic expression.

To implement our algorithm, we first search for the existence of a focus word or phrase in the transcribed user utterance. This process uses a conditional random field using a phrase-level dependency tree of the sentence aligned with part-of-speech tags as the input [25]. The algorithm labels each phrase with its likelihood to be the sentence focus. The most likely focus phrase, if its probability exceeds 0.5, is

Table 3 Response to statement cases

Case	Example
Question on focus word	U: "This game is very good to relax at the end of the day" S: "What kind of game?"
Partial repeat with rising tone	U: "I bought a new Kindle from Amazon" S: "A new Kindle?"
Question on predicate	U: 'I ate lunch late" S: "Where did you eat?"
Formulaic expression	U: "I do not think it is a good attitude to tackle such problems" S: "I see"

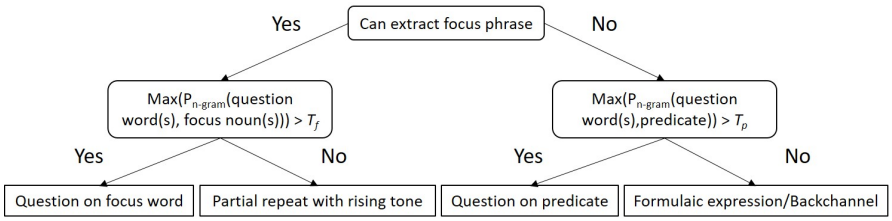

Fig. 3 Decision tree for statement-response.

assumed to be the focus. The resulting focus phrase is stripped so that only nouns are kept[1]. For example, the utterance "**The video game that has been released** is cool" would extract 'video game' as the focus noun.

We then decide the question marker to use as a response depending on whether a focus word can be found in the utterance. These transform an affirmative sentence into a question. Table 4 displays some examples of question words with and without a focus. We then compute the likelihood of the focus nouns associated with question words using an n-gram language model. N-gram probabilities are computed from the Balanced Corpus of Contemporary Written Japanese[2]. The corpus is made of 100 million words from books, magazines, newspapers and other texts. The models have been filtered so they only contain n-grams which include the question words defined above. The value of the maximum probability of the focus noun and question word combination is P_{max}. In the case where no focus could be extracted with a high enough confidence, we use an appropriate pattern based on the predicate. In this case, instead of the focus phrase, we compute sequences made of the utterance's main predicate and a set of complements containing question words. The best n-gram likelihood is also defined as P_{max}.

The second stage of the tree in Fig. 3 makes the decision into four leaves matching the four cases defined in Table 3. Each of those define a different pattern in the

[1] In Japanese, there are no articles such as 'a' or 'the'

[2] https://www.ninjal.ac.jp/english/products/bccwj/

response construction. T_f and T_p have been empirically fine tuned. Table 5 displays the conditions for generating each response pattern.

Table 4 Question words

With focus	Without focus
Which is/are	What kind of things
What kind of	When
When is/are	Where to
Where is/are	Where from
Who is/are	From who

Table 5 Response to statement methods

Case	Condition	Pattern
Question on focus word	$P_{max} \geq T_f$	question word(s) + focus noun(s) + "desu ka"
Partial repeat with rising tone	$P_{max} < T_f$	focus noun(s) + "desu ka"
Question on predicate	$P_{max} \geq T_p$	question word(s) + predicate + "no desu ka"
Formulaic expression	$P_{max} < T_p$	"So desu ka", "Tashikani", "Honto?"[3], etc.

3.3 Proactive initiator

As shown in Table 1, the proactive initiator takes part in several scenarios. Typical spoken dialogue systems are built with the intent of serving or reacting to the user speech inputs, while a situated system such as Erica continuously monitors its environment in search of cues about the intent of the user. This kind of interactive setup has been the focus of recent research work [18, 16, 20, 6, 5]. Erica uses an event detector to track the environment and generate discrete events. For example, we define three circular zones around Erica as her personal space (0-0.8m), social space (0.8-2.5m) and far space (2.5-10m). The system triggers events whenever a previously empty zone gets filled or when a crowded one is left empty. We also measure prolonged silences of fixed lengths.

Currently, we use the proactive initiator for three scenarios:

1. If a silence longer than two seconds has been detected in a question-answering dialogue, Erica will ask a follow-up question related to the most recent topic.

[3] "So desu ka": "I see", "Tashikani": "Sure", "Honto?": "Really?"

2. If a silence longer than five seconds has been detected, Erica starts a 'topic introduction' dialogue where she draws a random topic from the pool of available ones using a weighted distribution which is inversely proportional to the distance to the current topic in the word-embedding space.
3. When users enter or leave a social space, Erica greets or farewells them.

4 Evaluation and Discussion

The goal of our evaluation is to test whether our system can avoid making generic fallback utterances under uncertainty while providing a suitable answer. We first evaluate the statement response system independently. Then we evaluate if this system can reduce the number of fallback utterances. As we have no existing baseline, our methodology is to conduct an initial user study using only the question-answering system. We then collect the utterances from users and feed them into our updated system for comparison.

We evaluated the statement response component by extracting dialogue from a chatting corpus created for Project Next's NLP task[4]. This corpus is a collection of 1046 transcribed and annotated dialogs between a human and an automatic system. The corpus has been subjectively annotated by three annotators who judged the quality of the answers given by the annotated system as coherent, undecided or incoherent. We extracted 200 user statements from the corpus for which the response from the automated system had been judged as coherent.

All statements were input into the statement response system and two annotators judged if the response was categorized correctly according to the decision tree in Fig. 3. Precision and recall results are displayed in Table 6.

Table 6 Evaluation of statement response component

Category	Precision	Recall
Question on focus word	0.63 (24/38)	0.46 (24/52)
Partial repeat with rising tone	0.72 (63/87)	0.86 (63/73)
Question on predicate	0.14 (3/21)	0.30 (3/10)
Formulaic expression	0.94 (51/54)	0.78 (51/65)

Our results showed that the decision tree correctly selected the appropriate category in the majority of cases. The difference between the high performance of the formulaic expression and the question on predicate shows that the decision threshold in the case of no focus word could be fine-tuned to improve the overall performance.

[4] https://sites.googles.com/site/dialoguebreakdowndetection/chat-dialogue-corpus

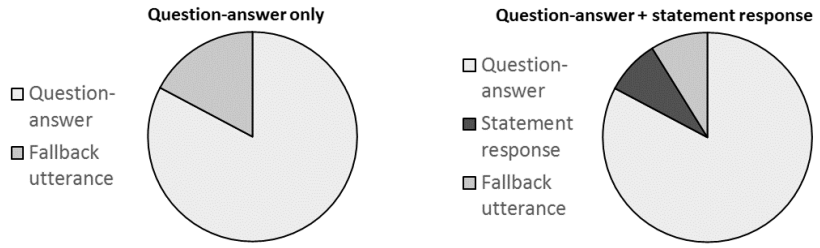

Fig. 4 Proportion of system turns answered by a component in the experiment (left) and the updated system including statement response (right).

We then tested whether the integration of statement response into Erica's dialogue system reduced the number of fallback utterances. The initial user study consisted of 22 participants who were asked to interact with Erica by asking questions to her from a list of 30 topics such as Erica's hobbies and favorite animals. They could speak freely and the system would either answer their question or provide a fallback utterance, such as "Huh?" or "I cannot answer that".

Users interacted with Erica for 361 seconds on average ($sd = 131$ seconds) with a total interaction lasting on average 21.6 turns ($sd = 7.8$ turns). From 226 user turns, 187 were answered correctly by Erica and 39 were responded to with fallback utterances. Users also subjectively rated their interaction using a modified Godspeed questionnaire [2]. This questionnaire measured Erica's perceived intelligence, animacy and likeability as a summation of factors which were measured in 5-point Likert scales. Participants rated Erica's intelligence on average as 16.8 (maximum of 25), animacy as 8.8 (maximum of 15), and likeability as 18.4 (maximum of 25).

We then fed all utterances into our updated system which included the statement response component. User utterances which produced a fallback response were now handled by the statement response system. Out of 39 utterances, 19 were handled by the statement response system, while 20 could not be handled and so again reverted to a fallback response. This result showed that around half the utterances could be handled with the addition of a statement response component, as shown in Fig. 4. The dialogues produced by the statement response system were generally coherent with the correct focus word found.

5 Conclusion

Our dialogue system for Erica combines different approaches to build and maintain a conversation. The knowledge and models used to cover a wide range of topics and roles are designed to improve the system's flexibility. We plan on improving the components using data collected through Wizard-of-Oz experiments.

While the question-answering system is simplistic, it can yield control to other components when uncertainty arises. The statement response mechanism helps to

continue the conversation and increase the user's belief that Erica is attentive to her conversational partner. In the future we also aim to evaluate Erica's proactive behavior and handle errors in speech recognition.

Our experiment demonstrated that a two-layered decision approach handles interaction according to simple top-level rules. We obtained some promising results with our statement response system and intend to improve it future prototypes. Other ongoing research focuses on learning the component selection process based on data. The main challenge in this architecture is determining which component should handle the conversation, which will be addressed in future work.

6 Acknowledgements

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant Number JPMJER1401), Japan.

References

1. Banchs, R., Li, H.: IRIS: a chat-oriented dialogue system based on the vector space model. In: Annual Meeting of the Association for Computational Linguistics, July, pp. 37–42 (2012)
2. Bartneck, C., Kulić, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics **1**(1), 71–81 (2009)
3. Becker-Asano, C., Ogawa, K., Nishio, S., Ishiguro, H.: Exploring the uncanny valley with Geminoid HI-1 in a real-world application. In: Proceedings of IADIS International conference interfaces and human computer interaction, pp. 121–128 (2010)
4. Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Maat, M., Mckeown, G., Pammi, S., Pantic, M., Pelachaud, C., De Sevin, E., Valstar, M., Wollmer, M., Shroder, M., Schuller, B.: Builduing Autonomous Sensitive Artificial Listeners. IEEE Transactions on Affective Computing **3**(2), 165–183 (2012)
5. Bohus, D., Horvitz, E.: Managing Human-Robot Engagement with Forecasts and... um... Hesitations. In: International Conference on Multimodal Interaction, pp. 2–9 (2014)
6. Bohus, D., Kamar, E., Horvitz, E.: Towards Situated Collaboration. In: NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data, pp. 13–14 (2012)
7. Den, Y., Yoshida, N., Takanashi, K., Koiso, H.: Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In: 2011 International Conference on Speech Database and Assessments (COCOSDA), pp. 168–173. IEEE (2011)
8. DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., Morency, L.p.: SimSensei Kiosk : A Virtual Human Interviewer for Healthcare Decision Support. In: International Conference on Autonomous Agents and Multi-Agent Systems, 1, pp. 1061–1068 (2014)
9. Ha, E.Y., Mitchell, C.M., Boyer, K.E., Lester, J.C.: Learning Dialogue Management Models for Task-Oriented Dialogue with Parallel Dialogue and Task Streams. In: SIGdial Meeting on Discourse and Dialogue, August, pp. 204–213 (2013)
10. Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., Matsuo, Y.: Towards an open-domain conversational system fully based

on natural language processing. In: International Conference on Computational Linguistics, pp. 928–939 (2014). URL http://www.aclweb.org/anthology/C14-1088
11. Higashinaka, R., Isozaki, H.: Corpus-based Question Answering for why -Questions. In: Interantional Joint conference on Natural Language Processing, pp. 418–425 (2008)
12. Higashinaka, R., Nakano, M., Aikawa, K.: Corpus-based discourse understanding in spoken dialogue systems. In: Annual Meeting on Association for Computational Linguistics, vol. 1, pp. 240–247 (2003). DOI 10.3115/1075096.1075127
13. Higashinaka, R., Prasad, R., Walker, M.A.: Learning to generate naturalistic utterances using reviews in spoken dialogue systems. In: International Conference on Computational Linguistics, July, pp. 265–272 (2006). DOI 10.3115/1220175.1220209
14. Higashinaka, R., Sudoh, K., Nakano, M.: Incorporating discourse features into confidence scoring of intention recognition results in spoken dialogue systems. Speech Communication **48**(3-4), 417–436 (2006). DOI 10.1016/j.specom.2005.06.011
15. Leuski, A., Traum, D.: NPCEditor: Creating virtual human dialogue using information retrieval techniques. AI Magazine **32**(2), 42–56 (2011)
16. Misu, T., Raux, A., Lane, I., Devassy, J., Gupta, R.: Situated multi-modal dialog system in vehicles. In: Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction, pp. 7–9 (2013)
17. Ogawa, K., Nishio, S., Koda, K., Balistreri, G., Watanabe, T., Ishiguro, H.: Exploring the natural reaction of young and aged person with telenoid in a real world. JACIII **15**(5), 592–597 (2011)
18. Pejsa, T., Bohus, D., Cohen, M.F., Saw, C.W., Mahoney, J., Horvitz, E.: Natural Communication about Uncertainties in Situated Interaction. In: International Conference on Multimodal Interaction, pp. 283–290 (2014). DOI 10.1145/2663204.2663249
19. Shitaoka, K., Tokuhisa, R., Yoshimura, T., Hoshino, H., Watanabe., N.: Active listening system for dialogue robot. In: JSAI SIG-SLUD Technical Report, vol. 58, pp. 61–66 (2010). (in Japanese)
20. Skantze, G., Hjalmarsson, A., Oertel, C.: Turn-taking, feedback and joint attention in situated human-robot interaction. Speech Communication **65**, 50–66 (2014)
21. Su, P.H., Gašić, M., Mrksic, N., Rojas-Barahona, L., Ultes, S., Vandyke, D., Wen, T.H., Young, S.: On-line Reward Learning for Policy Optimisation in Spoken Dialogue Systems. In: ACL (2016)
22. Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., Swartout, W.: Ada and Grace: Direct interaction with museum visitors. In: Intelligent Virtual Agents, pp. 245–251. Springer (2012)
23. Traum, D., Swartout, W., Gratch, J., Marsella, S.: A virtual human dialogue model for non-team interaction. In: Recent trends in discourse and dialogue, pp. 45–67. Springer (2008)
24. Varges, S., Quarteroni, S., Riccardi, G., Ivanov, A.V.: Investigating clarification strategies in a hybrid POMDP dialog manager. Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue pp. 213–216 (2010)
25. Yoshino, K., Kawahara, T.: Conversational system for information navigation based on POMDP with user focus tracking. Computer Speech and Language (2015)
26. Yumak, Z., Ren, J., Thalmann, N.M., Yuan, J.: Modelling multi-party interactions among virtual characters, robots, and humans. Presence: Teleoperators and Virtual Environments **23**(2), 172–190 (2014)