Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Toward Designing a Realistic Conversational System: A Survey

Awrad Mohammed Ali and Avelino J. Gonzalez
Department of Computer Science
University of Central Florida
Orlando, Florida, USA
awrad.emad@knights.ucf.edu, avelino.gonzalez@ucf.edu

Abstract

Over the past few years, the importance of having realistic conversational systems that satisfy the user needs have encouraged researchers to investigate new approaches for intelligent systems that are able to handle the required dialogues. However, designing a realistic conversational system that is able to understand the nuances of human conversation is not an easy task; thus, researchers face several challenges. To address these challenges and to help guide future research, we provide in this paper an overview of the most recent works in conversational systems. We classify the current models based on their functional similarities to address common features among the described systems. We also summarize the methods and approaches used by each system, state the systems limitations, and highlight their strengths.

Introduction

The need for good conversational systems has grown rapidly. As a result, researchers have sought to develop new systems that communicate with the user more naturally. This is important to keep a user satisfied and not seek alternative ways to satisfy his/her needs. However, this is not a trivial task. Researchers have classified existing conversational systems into two broad categories: 1) chat-oriented systems designed for entertainment purposes, such as the works by Banchs and Li (2012) and Sugiyama et al. (2013); 2) task-oriented systems that are designed to help the user accomplish specific tasks, such as making flight and/or restaurant reservations, or systems that provide information about specific topics, such as the works by Kim et al. (2007) and Woo and Kubota (2013). Chat-oriented systems have been considered more challenging than task-oriented systems because the former require a broad range of utterances to cover the user inputs, while task oriented systems only need a limited number of utterances given that the task is usually limited to assisting a user on a specific topic.

Generally, the existing dialogue systems can be divided into two types: 1) Rule-based models that depend on predefined rules to match the user utterance with the responses in the system's database. These systems also use general responses from existing tables that can be applied to different utterances. However, systems based on this model need a large database to cover the diversity in the user utterances. 2) Example-based models that use dialogue examples that have been semantically indexed to the database, as the work by Bang et al. (2015). Systems using this model do not require a big database, as do systems using rule-based models. However, they can sometimes provide meaningless responses because it is impossible to cover all possible patterns that appear in real human utterances. To produce reasonable responses, researchers introduce several methods, including a combination of different models, where each model is responsible for providing information about a specific task, for example the works proposed by Banchs et al. (2013) and Planells et al. (2013); and use memory to remember the user's previous tasks/dialogue, etc. Hence, this paper seeks to provide a comprehensive review about such works and address the limitations in the current research.

Descriptive Techniques

We address common features among the discussed works and classify them into several groups: heterogeneous systems, multi-models systems, memory-based models, models that use machine learning, and models that are designed to handle out-of-domain responses.

Heterogeneous Systems

Researchers have introduced multifunctional systems to build effective applications that can assist users in more than one task using the same interface. Planells et al. (2013) introduce a multi-domain dialog system that combines three pre-existing dialog systems: personal calendar, sporting event booking and weather services. These sub-systems are independent from each other, thus the dialog manager communicates with them using a fake user. For example, to close the current domain and activate another one, the dialog manager sends a fake massage to the active domain implying that the conversation is finished. The system operates by making only one sub-system active at a time, which could result in losing some important information regarding the dialog. To tackle this issue, the system has a context register that stores information related to multiple domains (i.e., time, dates, and places) and adds them to the user dialog. The system's importance come from its ability to generalized to include more sub-systems without needing to change its underlying

structure because each system thinks it is the only one in the architecture.

Testing indicated that the system's accuracy decreased as more tasks become involved in the dialog. More specifically, the accuracy decreases from 94.3% (for one task) to 76% (for all the three tasks).

In the same year, Banchs et al. (2013) present AIDA (Artificial Intelligent Dialogue Agent), which has six different dialogue engines, each of them is responsible for answering questions related to a specific topic. The appropriate task and engine are selected mainly by the user intention interface model, while the task selection and domain decision are based on three different sources of information: 1) the user utterances, including semantics, and feature extracted from those utterances. 2) Active engine information (only one is active at a time). 3) System expectation, including most recent history of the user-system interaction, the profile history of that user, and the task hierarchy.

Unfortunately, it is hard to evaluate the performance of this work because no results were reported in the original paper. The authors only provide an example of AIDA interacting with a user.

DHaro et al. (2014) introduced a multifunctional conversational system called CLARA that uses a natural language search mechanism to combine two applications: provide information about a conference, and a tourist guide agent. The user interacts with the system using a mobile application, which receives the user queries and passes them to the server. The server communicates with the search modules in the system architecture to provide information to the user based on different resources, including databases, dictionaries, and models.

Tests measured the usage statistics during the conference, including the number of users, number of questions related to papers or tourism, etc. The authors reported that the system failed to answer about half of the users' queries because 75% of the questions were out-of-domain.

The common feature among AIDA (2013), Planells et al. (2013), and CLARA (2014) is that at each dialogue turn, there is only one active sub-system. However, activating one system at a time could have both positive and negative effects in the system performance. It is positive because handling one topic is generally easier than handling more topics at the same time. Yet, this can result in producing wrong answers when the user asks questions that fit more than one model.

Additionally, all the previously discussed models have simple dialogue managers because the systems are task-oriented, which require the user to have limited number of utterances that he/she can use. Hence, these systems do not require large databases.

Multi Models Systems

In this section, we describe works that use more than one method to generate the system's responses. One recent examples is a chatbot presented by Nio et al. (2013). The authors compare and contrast the performance of a chatting system using statistical machine translation model (SMT) with its performance using a combination of two example-based dialogue manager EBDM methods: syntactic-semantic similarity retrieval, and TF-IDF[1] based cosine similarity retrieval. The benefit of combining these models is to solve the limitations that occur when each model is used individually. This is because EBDM usually provides proper responses based on dialogue examples that are semantically interoperated to a database, and SMT can generate related responses to the user input, even if it has not been trained on similar responses (Nio et al., 2013).

For evaluation, the authors showed a bar graph that reflects the improvement of TF-IDF cosine similarity metrics after applying semantic similarity filter. Using this filter not only improves the system performance but also minimizes the time required to provide responses because it reduces the examples in the training set (Nio et al., 2013).

The authors reported that the work was evaluated both objectively and subjectively. The subjective evaluation shows that using each method separately almost perform as good as combining them and sometimes even slightly better. Therefore, combining both models did not result in improving the results.

Another system known as SARA was introduced by Niculescu et al. (2014) as a multi-modal dialogue system that provides touristic information as a mobile application. SARA's dialogue manager consists of two different strategies to determine system responses: a rule-based approach and an example-based approach. The most similar response to the user input is selected based on the cosine similarities between the two vectors, and by using TF-IDF weights (Niculescu et al., 2014).

This work also presents an interesting study on how to handle out-of-domain queries. This study implies that when the system receives such queries, it asks the user to provide more information related to that topic. Doing this enables the system to improve its database by adding new information.

SARA was evaluated across five different scenarios, ranging from asking about a specific place in the city to providing general information. To perform the evaluation, 10 test subjects were asked to interact with SARA using three scenarios from the five available ones. Later, evaluators provide feedback about the system responses in terms of usability, reliability, and functionality by using Likert scale statements. The highest percentage achieved was agreeing for all the three scenarios. However, the authors reported that a noticeable percentage of people disagreed with the system because of mistakes in describing directions and venues.

Later, Shibata, Egashira, and Kurohashi (2014) present a chat-like conversational system that has multiple reply generating modules. Each module has a specialized action, such as selecting a sentence from a Web news source, question-answering, finding a definition in Wiki pages, etc. These modules are independent from each other, which makes it easier to modify a module or add another one. The system selects the response based on the user input and the dialogue history.

[1]TF-IDF stands for Term Frequency-Inverse Document Frequency, a numerical measure that reflects the importance of a word in a document or corpus.

Seven examiners were asked to evaluate the system performance by providing feedback that indicate their level of satisfaction about the system responses. Most of the evaluators reported that the system performance had improved as more dialogues are involved in the conversation, which reveals that the system had learned some dialogue strategy.

Morbini et al. (2014) created a flexible multi-initiative dialogue manager known as FLoReS (Forward Looking, Reward Seeking) that has a set of operators. These operators are responsible for controlling the dialogue flow. Each contains a sub-dialogue structure that is represented by a tree of system/user action, and the resulting state after taking that particular action. The system actions can be one of the following: 1) an answer to the user input; 2) an update to the information in the database; or 3) a command to send an event to the dialogue manager. For each state in the sub-dialogue there is a reward to reach that state. Based on the received reward, the system decides which operator to select.

Using rewards to decide the system's behavior is a new approach, but it puts pressure on the system designer because he needs to determine the reward for each operator. Thus, it would be interesting to have a system that is able to decide the rewards based on its observation.

There were no results reported for the FLoReS system. However, Morbini et al. (2014) mentioned that users' feedback indicates their satisfaction of the system performance.

Generally, using multiple sub-systems can help generate relevant responses to the user input because each sub-system responds to its relevant topic. Moreover, in all the proposed models, each sub-system generates a response but only the one that has the highest score is chosen to produce the system feedback. As we discussed, each system calculates the relevant score differently. For example, to decide the best score Nio et al. (2013) use statistical machine translation and example-based dialogue manager, while SARA uses cosine-similarity. Yet, the system proposed by Shibata, Egashira, and Kurohashi (2014) did not calculate a score to decide the best response; instead, it calculates a reward to learn the best response based on the dialogue history and the user input.

The system proposed by Nio et al. (2013) has an advantage over the other models because it filters the data before using them to generate the system output.

Memory-Based Systems

In this section we discuss the effect of using memory in conversational systems by discussing recent approaches that use memory to save previous dialogue history and/or user profiles in the process of generating the system utterances.

Banchs and Li (2012) introduced IRIS (Informal Response Interactive System), a chat-oriented dialogue system that learns new concepts from users, and semantically relates them to its previous knowledge. The system saves profiles of previous users to recall previous conversations, and use them to chat with the user. IRIS considers user feedback in improving its future responses by employing a mechanism that allows the user to rate the system's responses.

No statistical results were reported about IRIS performance. The authors only provided some dialogue examples that reflect where IRIS is performing well and when it fails to produce reasonable answers.

Kim et al. (2014) presented a spoken dialog system that uses long-term memory to save the user's previous utterances and use them later as part of the system responses. This is done by collecting the user facts in term of triples (arg_1, relation, arg_2), and save them in memory. Hence, the system uses natural language processing tools, such as part of speech taggers (POS), dialogue act classifiers[2], and knowledge extractors to process the system input to extract the triples. The same triples are extracted for the system response according to a method called triple substitution. The triples extracted from the user and the system are matched by changing the system arg_1 or arg_2 with the corresponding arg_1 or arg_2 from the user utterances. Moreover, the system keeps track of the user interests by distinguishing between long, and short-term interests by defining a forgetting model. To insure the relevance of its provided responses, the system calculates a "relevance score" between the examples in the database, and the system response using statistical information. The system selects the response with the highest score.

The system was tested by measuring the ratio of reasonable responses among three variances of the system: baseline system that uses a score based on the similarity of the input and output data; a system that uses relevance score only; and a system that uses both memory and relevance score. Results improved significantly from the base line score of 57% to 75% when relevance score measurement was used. However, adding the memory lowers the previous score to 74%.

Bang et al. (2015) introduced a chat-oriented dialogue system that combines EBDM with the personalized long-term memory proposed by Kim et al. (2014). Additionally, this system also uses three features: 1) POS-tag to match the sentences; 2) named entity (NE) types and values to search for the appropriate response; 3) back-off model to provide responses to unmatched user's sentences with the examples in the database.

Dialogue Act (DA), and POS-tagging are used to match the user input to the corresponding examples in the database.

The benefit of using DA is to reduce the search space because the system searches only the examples in the database that have the same DA as the user input.

The work is evaluated by using different combination of its components, i.e., baseline that uses simple lexical similarity to find similar examples; a system that uses POS; a system that uses NE; and a system that uses both POS and NE. The results are also compared against that of ALICE (Wallace, 2004). Eight users were asked to interact with all the variations of the system by rating the system performance using a scale from 1 to 5. The results reflect that the system that uses both NE and POS has slightly higher rating (e.g., ALICE score was 3.4, while the system with both NE and POS score was 3.7).

Even though using memory showed some improvement in

[2]Dialogue act (DA): is an expression that denoted user intention, such as greetings, confirmation, Wh-questions, etc.

producing relevant responces, it did not prevent the systems by Kim et al. (2014) and IRIS by Banchs and Li (2012) from producing meaningless results. This is because of noisy examples in the database. Thus, filtering the data is important to eliminate this problem. Moreover, the system by Banchs and Li (2012) suffers from another problem: not being able to maintain consistency with its previous answers.

The system by Banchs and Li (2012) has an advantage over the systems by Kim et al. (2014) and Banchs and Li (2012) by having a back-off model to provide answers when no matches are found with the user input.

Models that Use Machine Learning

In this section, we highlight recent works in conversational systems where machine learning plays an important role in their architecture. Machine learning has been used widely for estimating the next user/system action and the transition state. This is mostly applicable to task-oriented models, as the transitions and actions are considered deterministically. One example is a work presented by Lison (2013). His system uses model-based Bayesian reinforcement learning to estimate transition models for dialogue management. The idea is to use Partially Observable Markov Decision Process (POMDP) to teach the system which action to take by interacting repeatedly with the user. In POMDP, the current state is inferred from the observation of the agent, thus after each action and sequence of observations, the agent updates its belief about the state.

Like any other model that uses reinforcement learning, the agent seeks the actions that maximize its cumulative reward to find the optimal policy. For the transition model, the system has two different approaches: standard multinomial distribution, where the parameters are encoded using the Dirichlet distribution; and probabilistic rules to reduce the number of parameters needed by capturing the domain structure in a compact view.

This work was evaluated using a human-robot interaction scenario. The experiment was applied to a robot that was asked to perform some tasks, both verbally and virtually. The user utterances are limited to 16 predefined dialogue acts and the robot has a limited number of actions, including physical and conversational actions.

The evaluation was performed on both multinomial distribution and on probabilistic rules. Both models show improvements in estimating the transition during interaction with the simulated user. However, the probabilistic distribution model converged faster than the model with multinomial distribution because it was better able to capture the domain structure with limited number of parameters.

Machine learning has other applications, such as predicting the dialogue acts for future responses, and the current topic. Thus, Yoshimura (2014) proposed a conversational system that generates casual responses using large-scale data. In order to generate the system responses, the system needs to understand and analyze the aim and the context of the user utterances.

This system generates the utterances using data extracted from the web. The extracted data are in form of nouns and their corresponding predicates (e.g., predicate (eat), noun (bread)). The nouns and their predicates are extracted based on common human knowledge.

It is difficult to tell how well this project performs because no results are reported. However, our purpose in discussing this work here is to provide a different use of machine learning in conversational systems.

Machine learning in the form of reinforcement learning has been applied by Shibata, Egashira, and Kurohashi (2014) (discussed previously in Multi Models Systems section) to learn the best strategy that the system can follow for future responses.

Through our discussion of the works presented by Lison (2013), Yoshimura (2014), and Shibata, Egashira, and Kurohashi (2014), we have seen that machine learning algorithms have been applied to predict future actions, tags, topic, and transitions. The ability to predict future events can help improve the system performance because the prediction is based on previous events.

Systems Designed to Avoid Out of Domain Responses

One of the problems that many conversational systems face is how to handle out-of-domain responses. Therefore, in this section we discuss three approaches used to cope with this problem.

Data Filtering The term ”data filtering” has been presented earlier in our discussion of the work by Nio et al. (2013). Additionally, Yoshino, Mori, and Kawahara (2011) introduce a spoken dialogue system that also uses data filtering. Filtering the data is essential here because this system uses information extracted from the Web to generate its responses, and the Web contains noisy information that needs to be excluded. The extracted information is represented by a predicate-argument (P-A) structure, and is extracted from the user input as well. However, to eliminate producing irrelevant responses and to ensure that the dataset contain useful information, the authors use TF-IDF and Naïve Bayes model to measure the importance of a word in a given domain or topic.

The results reflect that the system was able to match the answers between the Web and the user input for only 30% of the cases and it did not provide any answers with a percentage of 68%. These results are even worse when the input to the system uses a speech recognition system as the correct answers dropped to 19.4% and the system did not provide responses for 79% of the cases. However, using back-off models to produce partial matching improves the system results by increasing the correct responses to 66.2%.

Related Words Extraction Researchers have found another way to identify out-of-domain responses by relating the system responses to the topic extracted from previous dialogue turns. Based on this idea, Sugiyama et al. (2013) proposed an open domain conversational system that uses template filling of the most relevant words from the user utterances and by using related words extracted from Twitter using web-scale dependency structures to generate its answers.

Sugiyama et al. (2013) realized that using a template-filling approach did not generate appropriate responses in some cases. Therefore, they used another utterance generation model based on dialogue acts. This model predicts the dialogue act for the system response based on the user's dialogue act. The system uses DA to generate its response when the predicted dialogue act is greeting, filler, sympathy, non-sympathy, or confirmation (cases where template-filling do not produce answers).

The system evaluation is performed by ten users who gave scores based on the system performance for six different models where two of them are Retrieval-Self and Retrieval-Reply models (Sugiyama et al., 2013). The other models are using either all the extracted topics (from noun and predicates) or either one of them. The proposed system (that uses all the topic extraction methods) performed slightly better than other models when the average score across all the evaluation items was 4.0 while the average scores for retrieval-self and retrieval reply were 2.9 and 3.8 respectively.

The system performance indicates that users were somehow satisfied with the system responses; however, the users reported that the system fails to produce answers related to the system itself.

Higashinaka et al. (2014) presented an open domain conversational system that is fully based on existing natural language processing (NLP) modules. This system considers intention, topic, and content of the sentence when generating the system responses. To identify the intention, the system uses a pre-existing dialogue act estimation module, and question-type classification module. This system considers the noun phrase (NP) to be the topic, where it is extracted using a conditional random field module. Predicates and their arguments are used to identify the content of the sentence.

The system performance was compared to the performance of a rule-based model and a retrieval-based model. Additionally, four combinations of the system were shown, in which at each case the authors either enable one model or more. Here we discuss the results of the system that has all the models enabled because there is not much difference between these results and the other variants of the system. Thirty users were asked to answer eight questions related to the system performance using a 7-point Likert scale. We calculated the average score given for each system across the eight questions to discuss the results. The system performed almost as well as the rule-based model, as the average score given for the system was 3.3 and for the rule-based model was 3.7. Retrieval had the lowest score, 2.7.

Both systems by Sugiyama et al. (2013) and Higashinaka et al. (2014) used topic extraction from user utterances to ensure having coherent responses; however, each system used a different part of speech to extract the potential topic. As we mentioned, Sugiyama et al. (2013) extracted the topic from proper nouns, common nouns, or predicates while, the system by Higashinaka et al. (2014) used only noun phrases to determine the topic.

The system by Higashinaka et al. (2014) used several modules to generate its answers, which makes it less likely to have the problem reported by Sugiyama et al. (2013) of not being able to generate responses in many cases.

Context-Based System Out-of-domain responses can arise when the system misunderstands the user utterance, more specifically when the user communicates with the system using an Automatic Speech Recognizer (ASR) since most ASR systems have high error rates. Thus, Hung and Gonzalez (2013) introduce CONCUR, a conversational system that provides responses based on understanding the context (speaker intent) of the conversation instead of trying to understand the syntax of the user utterances. Doing this eliminates the risk associated with understanding the complete utterances.

Users were asked to evaluate the system performance by providing scores on a scale from 1 to 7. The users are asked to evaluate the naturalness and the usefulness of the system. CONCUR achieved marks of 4.1 and 4.5 for the naturalness and usefulness respectively. By comparing these scores with the scores of other peer systems, Amani and Hassan (Gandhe et al., 2009) , where Amani achieved 3.0 and 3.2 for the naturalness and the usefulness, and Hassan scores 3.5 and 4.0. Clearly, the proposed CONCUR system was able to achieve better scores.

Another set of results were shown to demonstrate CONCUR ability to achieve high usefulness score of 60.5%, when the average ASR word error rate (WER) was high (58.5%). These scores are compared to the scores of another system known as Virtual Kyoto agent, in which it gains 61.4% with WER of 29.4%. It is clear that CONCUR system usefulness score is almost similar to that of Virtual Kyoto agent that has a lower word error rate.

Challenges and Open Questions

Throughout our discussion of the recent works in conversational systems, we identified several limitations in the current research that need to be addressed in future research. One of these problems is the limited ability of some systems to understand user utterances. This does not necessarily mean a problem related to the method itself but in many cases, a failure in speech recognition system is the main culprit. This results in producing meaningless responses regarding to the user input. However, Hung and Gonzalez (2013) elevate this problem by focusing on context meaning instead of the user syntax. On the other hand, the problem of having meaningless responses could be the result of including buggy data in the dataset used to generate the system responses. Such in the case of IRIS system by Banchs and Li (2012). Thus, filtering the data before using it as part of the system database should show significant improvement in the results, as we saw in the works by Yoshino, Mori, and Kawahara (2011) and Nio et al. (2013).

Additionally, the system performance could be improved further if researchers consider using memory to remember the user profile and the flow of the conversation. Doing this could eliminate out-of-domain responses and help link the current responses with the previous utterances in the dialogue. We had seen in our discussion that there are some works that use a combination of more than one method, but the current research is still missing a work that takes advantage of combining all the methods, i.e., use of memory,

filter the data, connect to the web to keep up-to-date information, and use human extracted information (i.e., topic, related words, etc.) to generate the system responses.

We also hope to see in future research better evaluation of the described systems to reflect better understanding of the systems' behaviors. A standard evaluation approach and benchmarks would be highly beneficial.

Summary and Final Remarks

In this paper, we have reviewed recent works in the area of conversational systems. We focused our discussion on how the described systems are able to understand the human utterances and generate their responses. We classified the works into five categories based on their functional similarities, including heterogeneous systems, in which the models were designed to accomplish more than one task; multi-models systems where the system's responses are generated based on multiple models; memory-based systems, where memory is included in the designed systems; models that use machine learning in their architecture; and finally models that are aimed to handle out-of-domain responses. As we discussed in previous sections, these systems generally work reasonably well and represent an improvement over the previous generations of conversational systems. Nevertheless, challenges and opportunities remain for future research in this exciting branch of artificial intelligent.

References

Banchs, R. E., and Li, H. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, 37–42. Association for Computational Linguistics.

Banchs, R. E.; Jiang, R.; Kim, S.; Niswar, A.; and Yeo, K. H. 2013. Aida: Artificial intelligent dialogue agent. In *Proceedings of the SIGDIAL 2013 Conference*, 145–147.

Bang, J.; Noh, H.; Kim, Y.; and Lee, G. G. 2015. Example-based chat-oriented dialogue system with personalized long-term memory. In *Big Data and Smart Computing (BigComp), 2015 International Conference on*, 238–243. IEEE.

DHaro, L. F.; Kim, S.; Yeo, K. H.; Jiang, R.; Niculescu, A. I.; Banchs, R. E.; and Li, H. 2014. Clara: a multifunctional virtual agent for conference support and touristic information.

Gandhe, S.; Whitman, N.; Traum, D.; and Artstein, R. 2009. An integrated authoring tool for tactical questioning dialogue systems. In *6th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 10–18.

Higashinaka, R.; Imamura, K.; Meguro, T.; Miyazaki, C.; Kobayashi, N.; Sugiyama, H.; Hirano, T.; Makino, T.; and Matsuo, Y. 2014. Towards an open domain conversational system fully based on natural language processing. In *Proceedings of the 25th International Conference on Computational Linguistics*, 928–939.

Hung, V. C., and Gonzalez, A. J. 2013. Context-centric speech-based human–computer interaction. *International Journal of Intelligent Systems* 28(10):1010–1037.

Kim, S.; Lee, C.; Jung, S.; and Lee, G. G. 2007. A spoken dialogue system for electronic program guide information access. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, 178–181. IEEE.

Kim, Y.; Bang, J.; Choi, J.; Ryu, S.; Koo, S.; and Lee, G. G. 2014. Acquisition and use of long-term memory for personalized dialog systems. In *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*. Springer. 78–87.

Lison, P. 2013. Model-based bayesian reinforcement learning for dialogue management. *arXiv preprint: 1304.1819*.

Morbini, F.; DeVault, D.; Sagae, K.; Gerten, J.; Nazarian, A.; and Traum, D. 2014. Flores: a forward looking, reward seeking, dialogue manager. In *Natural Interaction with Robots, Knowbots and Smartphones*. Springer. 313–325.

Niculescu, A. I.; Jiang, R.; Kim, S.; Yeo, K. H.; DHaro, L. F.; Niswar, A.; and Banchs, R. E. 2014. Sara: Singapores automated responsive assistant, a multimodal dialogue system for touristic information. In *Mobile Web Information Systems*. Springer. 153–164.

Nio, L.; Sakti, S.; Neubig, G.; Toda, T.; and Nakamura, S. 2013. Combination of example-based and smt-based approaches in a chat-oriented dialog system. *In Proceedings of ICE-ID*.

Planells, J.; Hurtado, L. F.; Segarra, E.; and Sanchis, E. 2013. A multi-domain dialog system to integrate heterogeneous spoken dialog systems. In *INTERSPEECH*, 1891–1895.

Shibata, T.; Egashira, Y.; and Kurohashi, S. 2014. Chat-like conversational system based on selection of reply generating module with reinforcement learning. In *Proceedings of the 5th International Workshop Series on Spoken Dialog Systems*, 124–129.

Sugiyama, H.; Meguro, T.; Higashinaka, R.; and Minami, Y. 2013. Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures. In *Proceedings SIGDIAL*, 334–338.

Wallace, R. S. 2004. *The anatomy of ALICE*. Artificial Intelligence Foundation Inc.

Woo, J., and Kubota, N. 2013. Conversation system based on computational intelligence for robot partner using smart phone. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, 2927–2932. IEEE.

Yoshimura, T. 2014. Casual conversation technology achieving natural dialog with computers. volume 15. NTT DOCOMO Technical Journal.

Yoshino, K.; Mori, S.; and Kawahara, T. 2011. Spoken dialogue system based on information extraction using similarity of predicate argument structures. In *Proceedings of the SIGDIAL 2011 Conference*, 59–66. Association for Computational Linguistics.