Towards a Method For Evaluating Naturalness in Conversational Dialog Systems

Victor Hung, Miguel Elvir, Avelino Gonzalez, Ronald DeMara
Intelligent Systems Laboratory
University of Central Florida
Orlando, Florida
victor@isl.ucf.edu, miguel.elvir@isl.ucf.edu, gonzalez@ucf.edu, demara@mail.ucf.edu

***Abstract*— The evaluation of conversational dialog systems has always remained a controversial topic, as it is unclear on how to quantitatively describe how well a conversation agent performs, or how much better one is over another. Furthermore, one of the hurdles in this quandary remains the definition of naturalness, as in how well a dialog system can maintain a natural conversation flow. This paper surveys existing evaluation practices and provides its own methodology to determine the effectiveness and naturalness of a dialog system. A prototype evaluation methodology for the LifeLike virtual avatar project is presented.**

***Keywords*— dialog systems, artificial intelligence, human-computer interaction, software evaluation**

I. Introduction

Chatbots, or interactive conversation agents, present a special challenge with respect to validation and verification. Specifically, the verification and validation of such software is not a process that solely relies on quantitative methods, as there remains a great deal of subjective evaluation involved in assessing their performance. Hence, the evaluation of chatbots remains a controversial topic, as there is no general method for judging how well a conversation agent performs, in both the relative and the absolute sense. In exploring this subject, a pivotal focus of this paper will be defining *naturalness*, as in how well a chatbot can maintain a natural conversation flow. This paper presents a survey of existing chatbot evaluation methods, as well as a definition for naturalness in relation to HCI applications.

This paper concerns the needs of a National Science Foundation (NSF) supported endeavor – the LifeLike virtual avatar project. LifeLike incorporates a conversational dialog system in its user interface. The prime directive of LifeLike is to provide expert decision support to its user base, while maintaining a sense of naturalness in its conversation-based human-computer interactions. Preliminary efforts in evaluating its dialog system have included both qualitative and quantitative measures. The objective of this paper is to investigate a proper methodology of chatbot evaluation for the purpose of validating the performance of LifeLike.

The remainder of this paper discusses the background technologies involved in chatbot evaluation, followed by a basic framework of the prototypical assessment system to be utilized by LifeLike.

II. Background

In order to empirically evaluate the naturalness of a dialog system's interaction with human users, we must first revisit the conceptual background underlying such applications. The proceeding section considers the background of issues in chatbot technology. This is accomplished by exploring the conclusions drawn by researchers whose applications reveal typical phenomena of naturalness and interaction.

A. *Early Intelligent Systems*

Early intelligent applications acted on declarative knowledge to process data. In these production systems, development of the learning framework relied to a large extent on explicitly defined rules with the purpose of assimilating new knowledge and exerting conflict resolution schemes. These models operate and maintain highly domain constrained knowledge bases whereby the user or client becomes the major recipient of the system's conclusion or hypothesis. Therefore, early production systems inherently provided immutable information retrieval processes, or a fixed context, with limited capacity to assess the validity of system output and modify its actions accordingly.

These intelligent agents simulated human performance in simple tasks by creating "goal-oriented and data-determined behavior" that relied on information processing and problem solving paradigms [11]. Declarative knowledge in the form of production rules governed the information retrieval and context selection phases [13]. Within this framework, context can be selected, matched against known scenarios, or traversed in predetermined directions by the agent's use of a fixed set of rules.

An example of early benefactors of these production systems took the form of chatbots that interact with a human operator via text and mimic human responses. One of the most successful and recent of these, ALICEbot by Wallace [12], could maintain short realistic conversations before digressing. Future implementations of ALICE type bots presented results from domain limited dialogs suggesting improved performance when incorporating user-initiated system corrections and several thousand response rules [12]. These applications, although appearing to maintain realism and a coherent shallow common sense, lack the ability to exploit symbolism in human understanding.

B. Naturalness in Dialog Systems

As the state of Automatic Speech Recognition (ASR) technology improves at recognizing human speech, increasingly sophisticated response systems will need to be developed that construe more natural dialogues with the human user and actively acquire new information from them [14]. Gurevych demonstrates, through an evaluation of the semantic coherence of ontology based speech recognition systems in [14], that a gap exists in recognition that is effectively and semantically coherent, partially due to the arbitrary nature of human speech and understanding of context [15].

Although at the present time spoken interaction may not be as efficient at accomplishing tasks as written interaction, Le Bigot et al. [15] suggest that such interaction promotes collaboration rather than placing emphasis on the task and its performance without regards to the dialogue quality. It can be inferred from their findings that this may be a result of both the lower informational density of speech and the elimination of essential terms for grounding 'shared knowledge' that occurs in human-computer speech interactions. As an example, consider the simple case of a computer prompting the user for the date on which he is arriving at a conference. The user's response may be as succinct as 'The twelfth', indicating a vague temporal sense. It is necessary, therefore, for intelligent speech applications to assert confirmation of any declarative knowledge it acquires throughout the interactive session in a manner consistent with these constraints.

Context retrieval experimentation in the form of meta-cognitive application development revealed that the acquisition of novel skills by an application can be facilitated by monitoring the state of knowledge, rationalizing goals, and implementing an adept instructional structure [16]. Intelligent tutoring systems by Pirolli [16] based on the ACT theories of knowledge suggest that some level of meta-cognition could improve a system's performance level at accomplishing its task. Extrapolating these findings to ASR knowledge frameworks would imply that the internal structure of the knowledge corpus and the system's awareness of its state and quality will directly impact the effectiveness of human-computer interaction. In particular, it is important that in extracting the relevant segments of a conversation the agent discover whether the new knowledge enriches the context of the interaction or whether it is detrimental to it.

A Knowledge Acquisition agent depends on the quality of the information received to identify the conversational domain [17]. An obvious impediment to obtaining contextually relevant data arises from imperfect transcripts from speech recognition. Semantic checks on the retrieved audio will hinder the system's interpretation of facts and its ability to validate context as seen in [14] and [15]. A simple experiment on this would show the input of an errorneously transcribed script into a chatbot and the effect on the bot's coherence. In the chatbot, the loss of information from transcription and structural organization becomes evident as the conversation progresses. While chatbots communicate directly with users via textual means, the conversation structure permits the information to maintain higher data density than transcripts of spoken communication [12]. Hence, our research may need to address the acquisition of sufficient spoken data to construct domain models. Retrieving subsumed themes from previous and current conversations imposes on the intelligent system the additional task of verifying the accuracy of its inferences and responses.

Per Schumaker, conversation length will be as imminent a factor in maintaining dialog quality as in chatbot applications [12]. Alice-based bots evince the need for assessing information quality on the part of the user and computer in order to quantitatively assess the relevance of new data to the current or emerging contexts. Such a principle falls within the metadata frameworks advocated by [18]. As can be seen from the results of Gurevych [15], the 'gold standard' for this would be the consensus of human judges with the system's interpretation of the domain.

C. Recent Advances and Integration of Realism

The copious research into artificial intelligence techniques since the inception of Eliza effected a lifting phenomena into chatbots. Evidence of such can be seen in the projection of cognitive, human behavior, and realism models in applications with chat-based roots. In this section, we chronologically demonstrate the direction of chatbot research. By implication, we also describe dialog systems and the work performed in advancing naturalness. As support, references to concepts supporting new theories will be provided.

Mateas [19] comparatively provides an overview of advances in chatbot related technologies for the late 1990's. Specifically, he demonstrates the initial departure, in this timeframe, from Eliza [8] style bots that employ sentence template matching. Instead, he highlights the increased importance of developing simple conversational memory. According to Mateas, several early systems capable of using conversational memory include multi-user dungeons (MUDs) such as the Julia project by Carnegie Mellon and Erin by Extempo. However, he notes several key differences between the conversational characteristics between these chatbots and more believable agents. Namely, interaction occurs in a reactive manner without pursuit of a goal by the bot. Moreover, these chatbots were intended to perform under a constrained version of the Turing test and only briefly fool a human.

In [20], Wlodzislaw et al. (2006) further expand on the naturalness restrictions evident in the template matching approach of Eliza-style programs. Briefly stated, earlier systems lacked domain expandability and could not fully exploit memory and reasoning components. Furthermore, they suggest that reliance on template matching can be associated to three key aspects of chatbots: focus on the Loebner prize, template-based AIML techniques, and the slow development of reasoning from natural language in dialogue systems. From [20], we learn that the development of cognitive modules and human interface realism for chatbot-like systems distinguishes avatars from Eliza- or ALICE-bot agents. As an example, they cite the use of ontologies, concept description vectors, semantic memory models, and CYC as tools that can serve to replace AIML templates and to increase the impression of understanding by the agent.

Works such as [20] and [21] reflect a shift towards enhanced immersive reality for dialog systems. It can be observed from these that a high emphasis on face-to-face avatar presentations and improvements in dialog evaluation will be focal points for future research. Traum and Rickel in [21] identify two considerations that present challenges to dialog management: multi-modal interaction and multi-party conversations. Becker and Wachsmuth [22] explore the representation and actuation of coherent emotional states in a virtual conversational agent. As an extension to the previous, [23] presents a model for sustainable conversation in a real-world application. Lars et al. [23] discuss several cognitive modules that increase the system's awareness of the human users and conversation topics. However, their system also relies on textual input in the same form as Eliza.

Some interest has been generated on the use of natural language processing for reasoning about human speech. However, several NLP applications may not be mature enough for implementation in bot-style applications. Furthermore, the tasks involved therein differ from those of natural language generation. For further exploration of methods for acquiring knowledge by agents, we direct the reader to [10]. Finally, a sample of reasoning from natural language processing techniques can be gleaned by analyzing the works referencing Lenat's Cyc, George Miller's WordNet, Berkely's FrameNet, etc.

From the approaches employed by the aforementioned bots, we perceive that an emphasis exists on developing goal oriented dialog systems that respond naturally. The principal efforts of this movement focus on creating more sophisticated interpretative conversational modules. Given the differences in techniques used to develop these bots, a need exists for generalizable metrics that evaluate the quality of a conversation in addition to the bot's performance.

The next section provides an overview of the issues considered to build an appropriate evaluation method for the LifeLike virtual avatar project. Furthermore, a prototypical system is proposed.

III. Approach

The development of LifeLike, as with any software creation, calls for a proper method of evaluating its performance. The challenging aspect of LifeLike, however, results from its identity as a vehicle of human behavior emulation. This means that the approach we will use for its evaluation process must incorporate elements of subjectivity from its human operators. This section discusses the duality of qualitative and quantitative aspects needed for chatbot evaluation.

A. Previous Approaches

Previous attempts at evaluating conversation agents all reflect a mix of quantitative and qualitative measures. Typically, subjective matters have involved a human user questionnaire. Semeraro et al [3] employ this technique for their bookstore chatbot. In the questionnaire, seven characteristics were appraised: impression, command, effectiveness, navigability, ability to learn, ability to aid, and comprehension. Users would assess their associated satisfaction for each of these metrics, ranging from 'Very Unsatisfied' to 'Very Satisfied.' Semeraro et al recognize the fact that this subjective evaluation does not provide statistically verified conclusiveness, but rather it serves as a general indicator of performance.

Shawar and Atwell [4] propose a universal chatbot evaluation system. They suggest three metrics, which were applied upon an ALICE-based Afrikaans conversation agent. The first metric concerns dialog efficiency, which deals with: atomic matching types, first word matching types, most significant matching types and no matching types. These matching methods establish how effectively a chatting agent can respond to user input. In their testing, Shawar and Atwell saw that first word matching and most significant matching were the most competent techniques. The second metric is the dialog quality metric, which qualitatively categorizes, by human judgment, a chatting agent's responses into three bins: reasonable, weird but understandable, and nonsensical. The final metric is users' satisfaction, which is also qualitatively measured. Feedback from the chatting software end-users is collected and used to directly evaluate the agent's performance.

Despite their efforts to establish a set of generic metrics, Shawar and Atwell [4] discourage the use of such a universal conversation agent evaluation mechanism. Instead, they conclude that the proper assessment of chatbots is the end result in how successfully it accomplishes its intended goals.

Evaluation of maintaining naturalness in a conversation similarly suffers from the same inherent problems of the general chatbot assessment system. Again, subjectivity plays a large role in judging the naturalness of a conversation. Rzepka et al [2] used a 1-to-10 scale for two metrics: a "naturalness degree," and a "will of continuing a conversation degree." In this study, human judges used these measures to evaluate a conversation agent's utterances. While their assessment system did not identify a concrete baseline for universal naturalness, they were able to make relative measurements of naturalness between different dialogue management approaches, such as comparing an ELIZA-based [8] manager with a WWW-based commonsense retrieval system.

Chatbot evaluation remains an open problem, especially because of its dependence on subjective assessment. Researchers use questionnaire-based methods to provide general insight on the effectiveness of their conversation agents. Similarly, measuring conversational naturalness also relies on user subjectivity. The major pitfall of these evaluation methods is their lack of quantitative universality, as no set of chatbot performance metrics has been defined. Nevertheless, current research has found success in using these techniques to make relative comparisons between conversation agents. Conversation agent evaluation, with emphasis on naturalness, plays a substantial role in appraising the performance of the work in this paper.

The remainder of this section gives a more in-depth treatment of the chatbot evaluation process, pointing out the primary factors that delineate the effectiveness of such dialog-based system software.

B. Chatbot Objectives

A dialog system, especially those of the *assistive* nature (as in LifeLike) proves its effectiveness under the light of two primary objectives: 1) dialog performance, and 2) task success. Each of these aims reflects different aspects of a human-computer conversation. Dialog performance relates to the experience of the interaction, while task success is concerned with the utility of the dialog exchange. Basically, these two objectives separately assess the effectiveness of the means (dialog performance) and the ends (task success).

The main goal of a dialog system is to achieve task success and dialog performance levels that are: 1) better than other dialog system solutions, and 2) similar to a human-to-human interaction. The latter stipulation defines the measure of naturalness, where a dialog system that can conduct a conversation that is similar to one between two people is considered natural. The next sub-section provides the metrics necessary to measure task success and dialog performance.

C. Evaluation Metrics

The evaluation system featured in this paper is derived from the PARAdigm for DIalogue System Evaluation (PARADISE) [1]. Table I depicts the structure of the objectives and their corresponding metrics within PARADISE. In this diagram, the master objective is user satisfaction, which is comprised of task success and dialog costs. Walker et al [1] further break down the dialog costs to efficiency measures and qualitative measures. These PARADISE-based objectives directly reflect the task success and dialog performance objectives mentioned in the previous section. The next sections discuss the metrics involved in task success and dialog costs.

TABLE I. CHATBOT METRICS

Metric	Type	Data Collection Method
Total elapsed time	Efficiency	Quantitative Analysis
Total number of user/system turns	Efficiency	Quantitative Analysis
Total number of system turns	Efficiency	Quantitative Analysis
Total number of turns per task	Efficiency	Quantitative Analysis
Total elapsed time per turn	Efficiency	Quantitative Analysis
Number of re-prompts	Qualitative	Quantitative Analysis
Number of user barge-ins	Qualitative	Quantitative Analysis
Number of inappropriate system responses	Qualitative	Quantitative Analysis
Concept Accuracy	Qualitative	Quantitative Analysis
Turn correction ratio	Qualitative	Quantitative Analysis
Ease of usage	Qualitative	Questionnaire
Clarity	Qualitative	Questionnaire
Naturalness	Qualitative	Questionnaire
Friendliness	Qualitative	Questionnaire
Robustness regarding misunderstandings	Qualitative	Questionnaire
Willingness to use system again	Qualitative	Questionnaire

TABLE II. ATTRIBUTE-VALUE CONFUSION MATRIX [1]

	Departure City			
DATA	***ATL***	***BOS***	***CLT***	***DEN***
ATL	*16*		*1*	
BOS	*1*	*20*	*1*	
CLT	*5*	*1*	*9*	*4*
DEN	*1*	*2*	*6*	*6*
SUM	23	23	17	10

D. Task Success

The tasks involved with a dialog system are of a multiple-goal nature. Thus, for any conversation, all of these goals must be recognized and satisfactorily serviced for the entire task to be considered successful. Conversations are modeled as a set of attribute-value pairs. Every user goal (and sub-goal) corresponds to an attribute, and the dialog agent's response to those goals represents a value.

As in PARADISE [1], an attribute-value matrix is created for both the expected response and the actual agent response in a conversation. A confusion matrix is produced to identify the discrepancies between the expected and actual attribute-value pairings. Table II gives an excerpted version of Walker et al's example attribute-value confusion matrix [1].

Walker et al present an attribute-value confusion matrix for a travel schedule system, with Departure and Arrival attribute-value pairings [1]. Table II gives a representative depiction of this matrix. Let us assume the question asked to the travel scheduling chatbot is, "Which city has a departure time of *X* o'clock?" The rows represent the *actual* responses from the agent, and the columns reflect the *expected* values.

In this matrix, there are four possible values for the departure city question. The value ATL was correctly identified 16 out of 23 times, while DEN was agreed upon 6 out of 10 times. This type of accuracy data may be extrapolated from the attribute-value confusion matrix. From this information, task success, κ is computed as the percentage of 'right' responses given by the agent.

E. Dialog Costs

Dialog performance is defined as a function of two types of dialog costs: efficiency and quality. Efficiency costs refer to the resource consumption needed to accomplish a single task or sub-task. These attributes can be measured in a solely quantitative manner. Qualitative costs measure the actual conversational content. These metrics may be recorded quantitatively or qualitatively. For qualitative assessments, users are given a Likert scale-based questionnaire following their interactions, providing feedback on the dialog system's naturalness, friendliness, etc. Walker et al [1], Stibler and Denny [5], Charfuelán et al [6], and Hassel and Hagen [7] provide some examples on suitable dialog costs. The Table I delineates the relevant cost metrics for this paper.

F. Performance Function

To evaluate the total effectiveness of a dialog system in relation to its task success, κ, and its dialog costs, c_i, the following PARADISE-based [1] performance function is used

$$Performance = (\alpha * \mathrm{N}(\kappa)) - \sum_{i=1}^{n} w_i * \mathrm{N}(c_i)$$

In this relationship, κ is weighted by α, and each w_i is a weight on c_i. The weight assignments are established in an arbitrary, yet meaningful manner. The function, N, uses a Z-score normalization process to balance out the effects of κ and c_i on the overall system performance.

This performance function allows for a normalized method of comparing two different dialog systems using the same conversational task goals. A dialog system will be considered performing in a natural manner when its performance level matches that of a human-to-human dialog.

The approach considerations presented in this section preludes into the proposed evaluation system prescribed for the LifeLike project. The following section provides a description of the system's prototype.

IV. Evaluation System Prototype

The purpose of establishing a chatbot evaluation system is to devise an experimentation infrastructure that collects data to support the idea of an improved human-computer interaction experience over other conversation agent systems. Furthermore, this interaction must reflect closely to the characteristics of a human-to-human exchange under the same situational premises. The quality of the interactive experience is judged using the previously mentioned metrics and task success assessment method, ultimately using the collaborative performance function to give a single measure of its effectiveness.

During the experimentation process, a single conversational scenario is employed on five different dialog systems. The first is the fully operational Context-Based Reasoning (CxBR)-based dialog system [10] featured in the LifeLike avatar project. The second dialog system is a crippled version of the first one, where the dynamic context-switching functionality is turned off. In a third experiment, an ELIZA-based [8] dialog system is tested. The fourth dialog system is modeled after an automated-phone operator, and the final system utilizes what is known as a Wizard of Oz (WOZ) experiment [9], where a human interlocutor replaces the normally machine-based agent.

In each experiment, the user is assigned five pre-specified goals to achieve during his or her dialog system interaction. A verbatim log of each conversation is retained for quantitative analysis and the user fills out a system quality questionnaire at the conclusion of the experiment. These data sources are used to compute the performance measure of the dialog system. Five different users will test each system. The user base will be selected under the assumption that cultural bias should not be a major factor when compiling results.

After executing all 25 trials (5 systems, 5 trials per system), the performance of each agent is compiled and evaluated for comparative analysis. Upon careful examination of these results, conclusions regarding the dialog system can be made. Future iterations of this experimentation process can be used for comparing later builds or enhanced versions of the same chatbot. This is analogous to "normal" software engineering practices, where unit testing is employed to verify that baseline functionalities are still intact between iterations.

V. Conclusion

This paper focuses on the inherent challenge of providing a proper evaluation process for conversation agent software. An overview of historical background technologies were presented to attest to the idea that such a problem is truly a challenge in the software engineering realm. We specifically sought out the validation requirements of the LifeLike virtual avatar to frame the chatbot evaluation problem in a real-world treatment. A proposed assessment methodology for LifeLike was presented, a prototypical framework derived from the PARADISE [1] infrastructure.

VI. Future Work

The LifeLike virtual avatar is still in its prototypical stage, which precludes the immediate requirement for a verification and validation process. Preliminary evaluations of the avatar software's prototypes have been made, with much of the aforementioned material taken into consideration. A formal treatment of the evaluation process has yet to be implemented; thus, any formidable results have yet to be prepared in publishable form.

As our work with LifeLike progresses, so will our need to provide a reliable evaluation system. Establishing such a method will allows us to better judge the evolutionary direction of our software, as well as any other chatbot software outside of LifeLike.

Acknowledgment

This research is supported by NSF Collaborative Research award 0703927.

References

[1] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: a framework for evaluating spoken dialogue agents," Proc. of the 8th conference on European chapter of the Association for Computational Linguistics, 1997, pp. 271-280.

[2] R. Rzepka, Y. Ge, and K. Araki, "Naturalness of an utterance based on the automatically retrieved commonsense," Proc. of Nineteenth IJCAI, 2005.

[3] G. Semeraro, H. H. K. Andersen, V. Andersen, P. Lops, and F. Abbattista, "Evaluation and validation of a conversational agent embodied in a bookstore," Universal Access: Theoretical Perspectives, Practice and Experience. Lecture Notes in Computer Science, 2615, 2003, pp. 360-371.

[4] B. A. Shawar, and E. Atwell, "Different measurements metrics to evaluate a chatbot system," Proc. of the 2nd Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing, 2007.

[5] Stibler, K., and Denny, J., "A three-tiered evaluation approach for interactive spoken dialogue systems," Proceedings of the first international conference on Human language technology research, 2001, pp. 1-5.

[6] Charfuelán, M., Gómez, L. H., López, C. E., and Hemsen, H., "A XML-based tool for evaluation of SLDS," Proceedings of the Third International Conference on Language Resources and Evaluation, 2002.

[7] Hassel, L., and Hagen, E., “Evaluation of a Dialogue System in an Automotive Environment,” 6th SIGdial Workshop on Discourse and Dialogue, 2005.

[8] J. Weizenbaum, “ELIZA-a computer program for the study of natural language communication between man and machine,” Communications of the ACM, 9(1), 1966, pp. 36-45.

[9] Hajdinjak, M. and Mihelic, F., “A Dialogue-management Evaluation Study,” Journal of Computing and Information Technology, 15(2), 2007, pp. 111-121.

[10] B. Stensrud, G. Barrett, V. Trinh, and A. Gonzalez, “Context-based reasoning: a revised specification,” FLAIRS Conference, 2004.

[11] d'Ydewalle, G., & Delhaye, P., “Artificial intelligence, knowledge extraction and the study of human intelligence,” International Social Science Journal , 1988, pp. 63-72.

[12] Schumaker, R. P., Liu, Y., Ginsburg, M., & Chen, H., “Evaluating mass knowledge acquisition using the ALICE chatterbot: The AZ-ALICE dialog system,” International Journal of Human-Computer Studies , 2006, pp. 1132-1140.

[13] Anderson, J.,“ACT - A Simple Theory of Complex Cognition. American Psychologist,” 1996, pp. 355-365.

[14] Le Bigot, L., Terrier, P., Amiel, V., Poulain, G., Jamet, E., & Rouet, J.-F., “Effect of modality on collaboration with a dialogue system,” International Journal of Human-Computer Studies , 2007, pp. 983-991.

[15] Gurevych, I., Malaka, R., Porzel, R., & Zorn, H., “Semantic coherence scoring using an ontology,” Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. Morristown: Association for Computational Linguistics, 2003, pp. 9-16.

[16] Pirolli, P., & Recker, M.. “Learning Strategies and Transfer in the Domain of Programming,” Cognition and Instruction, 1994, pp. 235-275.

[17] Potter, S., “A Survey of Knowledge Acquisition from Natural Language,” TMA of Knowledge Acquisition from Natural Language, 2001.

[18] Hahn, U., & Schnattinger, K., “An Empirical Evaluation of a System for Text Knowledge Acquisition,” EKAW '97: Proc. of the 10th European Workshop on Knowledge Acquisition, Modeling and Management, 1997, pp. 129-144.

[19] Mateas, M., “An Oz-Centriv Review of Interactive Drama and Believable Agents,” In Artificial Intelligence Today: Recent Trends and Developments, 1997, pp. 297-343.

[20] Wlodzislaw, J., Szymański, J., Sarnatowicz,T. “Towards Avatars with Artificial Minds: Role of Semantic Memory,” Journal of Ubiquitous Computing and Intelligence, 2006, 1(in print).

[21] Traum, D. & Rickel, J. “Embodied agents for multi-party dialogue in immersive virtual worlds,” Proc. of the first international joint conference on Autonomous agents and multiagent systems: part 2, 2002, pp. 766-773.

[22] Becker, C. & Wachsmuth, I., “Simulating the emotion dynamics of a multimodal conversational agent,” Proc. Tutorial and Research Workshop on Affective Dialogue Systems (ADS-04), 2004, pp. 154-165.

[23] Lars, G., Krämer, N., Wachsmuth, I., “A conversational agent as museum guide – Design and evaluaiton of a real-world application,” The 5th International Working Conference on Intelligent Virtual Agents, 2005, pp. 329-343.