Dialogues in Context: An Objective User-Oriented Evaluation Approach for Virtual Human Dialogue

Susan Robinson, Antonio Roque, David Traum
Institute for Creative Technologies, University of Southern California
13274 Fiji Way,
Marina Del Rey, CA
United States
E-mail: robinson@ict.usc.edu, roque@ ict.usc.edu, traum@ ict.usc.edu

Abstract

As conversational agents are now being developed to encounter more complex dialogue situations it is increasingly difficult to find satisfactory methods for evaluating these agents. Task-based measures are insufficient where there is no clearly defined task. While user-based evaluation methods may give a general sense of the quality of an agent's performance, they shed little light on the relative quality or success of specific features of dialogue that are necessary for system improvement. This paper examines current dialogue agent evaluation practices and motivates the need for a more detailed approach for defining and measuring the quality of dialogues between agent and user. We present a framework for evaluating the dialogue competence of artificial agents involved in complex and underspecified tasks when conversing with people. A multi-part coding scheme is proposed that provides a qualitative analysis of human utterances, and rates the appropriateness of the agent's responses to these utterances. The scheme is outlined, and then used to evaluate Staff Duty Officer Moleno, a virtual guide in Second Life.

1. Introduction

Virtual agents have made rapid progress in recent years, particularly in the complexity of dialogue situations they are designed to encounter, however the methodologies for evaluating these agents has largely lagged behind. Agents are most commonly evaluated by some variety of three methods: (1) objective or subjective indications of task success, (2) an objective, largely automated analysis of component performance and interaction features or (3) a subjective usability evaluation based on user feedback to surveys. The first method is very powerful when the system is designed to perform only a single simple task, such as retrieving a simple piece of information. However, when the tasks are complex and do not have simple mappings as to how successfully they have been performed, and when different users attempt different tasks, it can be difficult if not impossible to draw meaningful conclusions. Moreover, even where tasks can be defined and measured, the success of a task does not necessarily entail success of the dialogue – such measures are appropriate for viewing dialogue agents as tools to complete a task, but not for viewing them as conversational partners. Automated component methods are important for developers, in that they give a clear view on how well a component is performing relative to its design, but they lack in two respects: the results can be compared only to another component using a similar design, and the results only show how well the component is performing in its design function—not whether that approach is actually effective in the dialogues between agent and user. User-based evaluations are more useful for giving a sense of dialogue success but lack sufficient objectivity and detail to be of serious use for developmental evaluation, as they give little sense of what strategies are working where in a complex dialogue.

What we need for our purposes is a method of evaluation that gives specific feedback for directing improvement of agent dialogue at multiple phases of development, and ideally is largely transferrable with minimal modifications, to the evaluation of multiple agents with differing domains and functionality. While the perceived ease and speed of automated methods of evaluation make them increasingly popular, there is reason to be skeptical of the ultimate utility of such methods: If we already had the knowledge and ability to do a successfully detailed automated evaluation of the complex actions involved in dialogue using formal methods, we would then already have most of the knowledge required to build perfectly functional dialogue systems. What is often needed, especially at early stages of development, are methods of identifying and understanding the features of a dialogue from a more detailed linguistic perspective. The current work presents criteria toward an approach to the analysis of agent dialogue that meets these requirements. Our approach is based on a paired coding scheme, in which user utterances are tagged with two linked tags: the dialogue action of the utterance and an evaluation of the agent's quality of response. We use this method to evaluate Staff Duty Officer Moleno (Jan et al., 2009), a conversational agent currently active and publically accessible in the online world of Second Life.

The paper is organized as follows: Section 2 will discuss the key issues in dialogue evaluation in more detail, with particular reference to domain oriented conversational agents. Section 3 describes the annotation schemes proposed. Section 4 presents the evaluation of Moleno. Section 5 discusses in more detail how this approach addresses the problems raised in Section 2, and outlines considerations for future work.

2. Issues in Dialogue Evaluation

Different evaluation methodologies are formed by a complex relationship between the type of system or component being evaluated, the purpose or goals of the evaluation, and the perspective of the evaluation. Generally they consist of one or more of the three types of evaluation described in the previous section, which are evaluated from the perspective of the system or the end-user. In this section we review several approaches to try to go beyond these perspectives and gain a better sense of dialogue quality, itself.

The PARADISE method (Walker et al., 1997) took some promising steps toward a general evaluation methodology by bridging the gap between objective function and a sense of quality and by providing a method of comparative evaluation. Taking user satisfaction as the end goal, it allows one to test dialogue features against user-based evaluations and to see which parameters yield significant correlations. In principle this may give predictive results of the effect of system parameters on user satisfaction. In practice, however, it is not clear this approach gives much useful information on dialogue, as the 'dialogue quality' measures — both the objective function and the resulting correlation with user satisfaction, are so general as to be of questionable usefulness. For example, the results from testing three systems with PARADISE (Kamm et al., 1999) identified two factors that had a clear positive correlation with user satisfaction: task success and 'dialogue quality', which was correlated with Mean Recognition Score (a mean rating of concept recognition accuracy across the whole dialogue). Furthermore, conflicting results in different PARADISE studies suggest that even some of these general dialogue quality measures are not usefully comparable across domains. In a PARADISE-style evaluation of a task-oriented human-robot dialogue system (Foster et al., 2009), dialogue length was one of three factors that correlated with user ratings. Contrary to previous studies, however, it was increased dialogue length that correlated with higher user satisfaction.

A present shortcoming of this framework for dialogue evaluation, and with other approaches such as user modeling, is the focus on a basic tenant of user-based evaluation: that quality should only be determined by the users of a system (King, 2007). While this is a fairly straightforward statement when 'quality' refers to the usability and function of simple software applications designed with a particular purpose for end-users, it does not follow that users have more than a general sense of dialogue quality or have the capacity to make judgments specific enough to provide useful results to help define those dialogue features that contribute to quality. While the PARADISE method may be solid, its application is limited by our current definitions of dialogue features. For better understanding of dialogue performance, we need finer-grained performance measures of dialogue quality defined in objective terms.

If we want to compare the performance of various dialogue systems, we need an evaluation perspective that is meaningful that holds across all of those dialogue systems. Paek (2007) reviews various evaluation approaches, including PARADISE and concludes that none is sufficient for comparability across diverse systems, or that ultimately such a metric may be impossible. Given the various design strategies, tasks and user needs of different systems, it should be clear that the only commonality that holds across these systems is that they are all engaged in dialogue. Therefore the only perspective from which they may be evaluated in common is from the perspective of the dialogue itself as the object of investigation, and measured against the only dialogue standard we all hold in common — the standard of human performance.

As complex dialogue performance takes center stage with non-task-oriented conversational agents, it is necessary to redefine the appropriate perspective relating to the larger goal: to approach human-like communication. While this may be a far-fetched goal, given current understanding of human language, it is a resurging concept that is appearing in recent literature (Edlund et al., 2008; Holzapfel et al., 2008). Ultimately it does not matter if we are far away from that goal, or really if it can ever be considered fully met — the shift of target goal in itself is enough to lead to new possibilities. Dialogue quality, regardless of how we explicitly define it, is ultimately implicitly measured against the norms or expectations of human dialogue. What we gain by making this an explicit goal in evaluation, is a new perspective on evaluation methodology. Note that this is different from the Turing test (Turing, 1950), in which the goal is to be indistinguishable from a human via dialogue performance. The goal here is human-like appropriate dialogue behavior, not mimicry.

Surprisingly little evaluation to date has taken place at the actual dialogue level — that is, interactive dialogue itself as the subject of evaluative study. But as the field of conversational agents matures, this will become an increasingly vital perspective. A key barrier in this work is a method for directly assessing the quality of a dialogue, and moreover, an ability to measure it in a well-defined objective manner analogous to, say, WER. These difficulties have led researchers to skip directly to the familiar method of user evaluation. While those targeting holistic system quality judgments of the user as the primary goal may see this as relatively unproblematic (Möller & Ward 2008), the gap is increasingly recognized by researchers attempting to stretch previous boundaries of conversational capability in non-task oriented agents (Bernsen et al., 2007; Artstein et al., 2009).

Aside from task performance, there are evaluation measures that have been proposed that do have more useful detail specific to dialogue performance. The TRINDI ticklist (Bohlin et al., 1999) proposed an evaluation method in the form of 12 questions on specific interactive capabilities of the system formalism. This work has been used by a number of researchers to qualitatively evaluate dialogue systems and formalisms. The questions are useful in that they do describe the complexity of linguistic interaction that the system is capable of, however, they concern only broad capabilities of systems, not quantitative evaluation of how often the systems successfully manage such phenomena on real dialogue or how frequent and/or important the phenomena are in particular interactions.

Like our approach, others have used annotation methods applied to human-agent dialogues for evaluative purposes. Key definitions of the annotations, however, are oriented toward the perspective of the specific capacities or function of the agent being evaluated, however, rather than toward an objective or external definition of measure of dialogue quality. DATE (Walker & Passonneau, 2001) is closer to the sort of interactive detail we are after. It consists of a detailed dialogue act tagging scheme applied to the agent's utterances. While it can give a useful characterization of the agent's behavior, it is very specific to that agent's design and domain. Danieli and Gerbino (1995) applied an 'appropriateness' scheme in evaluation, but appropriateness was defined for their use in terms specific to the agent's domain functions.

3. Annotation Methods

With these goals in mind, we have developed two coding schemes. The first scheme characterizes the apparent dialogue action of the user's utterance, the second evaluates the agent's response to that action, judged in terms of appropriateness.

3.1. User Dialogue Action Description

The first scheme is based on Robinson et al. (2008), with some specific modifications for the current domain. The original scheme was developed empirically to examine topical user preferences in dialogue with a question-answering character and proposed a hierarchical scheme for user utterance categorization. Since that agent had numerous prompts to direct the user into topics he was familiar with, it was necessary to divide all user questions into initiating topic vs. responding to an agents prompt at the top level. Since the current goal is to characterize the user's dialogue actions, this distinction was dropped. In addition, domain-specific actions were added for the current agent, which will be discussed further in section 4.2.

Code		Description
D		*General Dialogue Functions*
	DG	Greeting
	DC	Closing
	DP	*Politeness*
C		*Critique (of agent or domain)*
	CM	*Critique of Agent*
	CD	*Critique of Domain (Army World)*
E		Exclamations, Emotive Expressions
H		*Hazing, Testing, Flaming*
	HH	Hazing, Testing
	HF	Flaming
F		Flirting- Playful Question or Offer
Q		*Information Requests*
	QD	*Domain Related Information*
	QM	*Information about Agent*
	QC	Communication Modality Options
	QO	Other Information Requests
R		*Requests*
	RH	Request for Help (Generic)
	RC	Request Clarification
	RM	*Request Motion of Agent*
	RD	*Domain Specific Request*
A		*Answers to Agent Utterances*
S		*Statements*
	SC	Continuing Own Utterance
	SS	*Initiating Social Topic*
	SD	*Domain Related Statements*
	SR	*Responsive Statements*
	SO	Initiating Other Topic
L		Utterance in another language
G		Incomprehensible to coder
O		Other Addressee

Table 1: Dialogue Action Scheme (Top 2 Levels).

The user action description scheme is hierarchical in the sense that it combines several layers of specificity. The top layer consists of generic dialogue acts, some specific to human-agent interaction, such as hazing the agent. The second layer further subcategorizes some of the top level acts, but is still fairly generic. The third layer further narrows the action into domain specific and subtler topical distinctions. For the discussion here, we are concerned with the top layers, as shown in Table 1. Tags and descriptions shown in italics indicate that the category is further subcategorized, so these top level tags were not actually used in our test annotation. The choice of which level to annotate is influenced by the distributions of utterances in the data, given a desire to have categories that include neither too little of the data to draw any meaningful conclusions or too much to do any meaningful analysis. In another domain, further sub-categorization may be required in some acts, while others may be annotated at the top level, or not at all. For example, an airline reservation system might require a very detailed sub-categorization of domain-related answers, but only the top level for other acts, assuming they occur very infrequently, if at all. Defining different flexible levels of specificity allows results from different domains to be compared and contrasted, while still maximizing the utility of annotation for a particular system and/or domain.

A common approach to dialogue act annotation is where the occurrence of an act is at least partially based on speaker intention (Traum, 2000). We speak instead of 'apparent action' to distinguish from these approaches. To evaluate the quality of the agent's dialogue, what the speaker 'meant' is actually irrelevant. Even human-human dialogue is imperfect and speakers sometimes fail to express their intentions in a manner that a hearer understands. Instead, the annotator is instructed to align with the agent and annotate from a hearer's perspective, with no look-ahead to interpret the user's utterances. This orients the judgments toward what a human understands, given that dialogue context (see (Edlund et al., 2008) for a similar perspective).

3.2. Evaluative Coding Scheme

The second scheme, for evaluative coding, is a modified version of one proposed in Traum et al. (2004), which evaluates agent utterances based on a sense of appropriateness. The agent's action in response to the user's utterance was rated with one of the six tags shown in Table 2.

Code	Value
3	Appropriate Response
NR3	No Response (Appropriate)
2	Partially Appropriate
RR	Request Repair
NR1	No Response (Inappropriate)
1	Inappropriate Response

Table 2: Evaluative Coding Scheme

Appropriateness is defined in objective terms by comparison with a human dialogue perspective, rather than using a subjective notion of the evaluator as to the upper boundary of the capabilities of the agent. Appropriateness is judged relative to the action already tagged using the scheme from Table 1. A 'partially appropriate' response may lack some coherence or relevance, but is considered adequate. A novel aspect of this scheme is that we also annotate agent silences in terms of quality. While in many simpler agent domains silence is indicative of error (an agent's failure to respond), there are many situations where silence is a good thing. An over-reactive agent may respond inappropriately to back-channeling, or respond inappropriately to utterances addressed to others in a multi-party situation. In addition, silence in human-human conversation is multi-functional, and can sometimes be taken as a form of response. Therefore, when we discuss the 'appropriateness of the agent's response' in objective dialogue terms, 'response' is defined in terms of human - human dialogue and includes the possibility of an overt action (utterance) or non-action (silence). In addition to the two schemes discussed, relevant agent utterances were also tagged with *I* indicating initiative utterances, to gauge the degree of Moleno's initiative in the dialogues.

4. Evaluation of Staff Duty Officer Moleno

This section presents a test of our annotation methods in the evaluation of dialogues with Staff Duty Officer Moleno. We give an overview of the domain and system design, describe additional features of the dialogue action scheme and discuss the resulting evaluation.

4.1. System Description

Staff duty officer Moleno is active in Second Life, a public online virtual world where users explore and interact through virtual avatars. Moleno is a roaming agent in the Army Welcome island, where visitors can explore, learn about the US army, and participate in several activities, such as a parachute jump, helicopter ride and quiz. Users can gain points from these activities and trade them for virtual prizes in a gift shop. Moleno's main function is to help users navigate the island and answer any questions they may have. He is also available to give tours of the island. These functions are not necessarily known to users when they arrive in this region, however, nor is the fact that he is an agent, rather than another human-controlled avatar. As with any area in second life, users can opt to do none of the above activities, and simply roam the space and chat with the avatars of others. Another factor affecting dialogues is that everyone in the realm is represented by a virtual avatar—users are biased toward assuming Moleno is, like them, controlled by a real person, until some behavior suggests otherwise. Thus users may approach Moleno in a much different way than they would interact with a virtual guide in a museum or other real world space.

Moleno communicates with users through two text based modalities, instant messaging (IM) and chat. IMs can be addressed to only one user, and have no range limit in virtual space. Chats have a limited spatial range approximating real world communication and may be received by any number of users within this proximity. Chat dialogues in this world are always potentially multi-party, which presents additional challenges for the agent. The agent's core response selection is implemented using the NPCEditor (Leuski & Traum, 2010), which includes a classifier using cross-language information retrieval techniques, and is trained with a set of input utterances mapped to a finite set of responses. Moleno also has additional features to manage multiparty dialogue situations: he keeps a user model of everyone he meets (which persists between sessions). The user model includes information on the time elapsed since his last interaction with that user, whether they are typing, and their location. In multiparty situations, Moleno delays responding to utterances he is uncertain about, only taking a turn after a certain time of no users typing (Jan et al., 2009).

Code		Description
QD		*Domain Related Information Requests*
	QDG	Army Island General Information
	QDL	Location/Navigational
	QDE	Events/Activity Specific
	QDJ	How do I join the army?
	QDU	About getting a uniform
	QDW	About getting weapons
	QDA	Other (real world) about US Army
	QDM	Meta Domain (about the domain, as sim)
	QDV	What's a virtual soldier?
QM		*Information about Agent*
	QMB	Biographical information
	QMP	Agent's preferences (favorite food, etc)
	QMJ	Job/ Purpose
	QMI	Immediate Experience (what you doing?)
	QMA	Agency (are you a real person?)
	QMK	Knowledge/skills
RD		*Domain-Specific Requests*
	RDX	Request teleport or accompaniment
	RDT	Request a tour of Island
	RDO	Misc domain requests
A		*Responses to Agent Offers*
	AXY	Accept teleport
	AXN	Reject teleport
	ATY	Accept tour
	ATN	Reject tour
	AON	Other rejections
SD		*Domain-Related Statements*
	SDP	Statement of domain preference
	SDN	Own navigation/location ("I'm lost")
	SDE	Experience of domain (I've seen that)
	SDB	Bugs/ problems encountered in domain

Table 3: Dialogue Action Scheme (Bottom Levels)

4.2. Domain Specific Annotations

In addition to the dialogue action tags discussed in section 3, there is a third, more detailed layer of annotations that were defined for this study. While most of these are domain-related, there are a few exceptions. We subdivided 'Politeness' into four categories: "Nice to meet you" and similar polite statements (DPS), "how are you?" (DPH), thanks (DPT) and apologies (DPA). The specific codes used for both categories of critique (agent critique and domain critique) distinguished positive from negative critique. Request Motion (RM) was a generic category for requesting deictic or temporal actions of the hearer. The actual tags used were: request to go away (RMG), request to 'come here' (RMH), request to wait, or standby (RMW), and request to continue with an action or dialogue turn (RMC). Finally, certain reactive statements on level 3 aren't necessarily domain-related. These include acknowledgement or generic feedback (back-channeling) statements (SRA), canceling a topic after the listener expresses confusion (SRC), and miscellaneous reactive comments to an utterance by the agent (more semantic content than a simple acknowledgement, but not functionally a critique) (SRM). Otherwise, the level 3 annotations are domain specific, and listed in Table 3.

We did an initial test of inter-coder reliability of the coding schemes between the first author, who created the annotation guidelines, and the other authors, who read the guidelines without prior discussion. Results are shown in table 4.

Scheme	Score	A1-A2	A1-A3	A2-A3
Full Action Codes	P(A)	0.619	0.695	0.562
	Kappa	0.613	0.691	0.555
Top Level Only	P(A)	0.848	0.800	0.771
	Kappa	0.836	0.787	0.756
Evaluative Codes	P(A)	0.748	0.725	0.806
	Kappa	0.734	0.710	0.793

Table 4: Agreement and Kappa

The Kappa value was moderate for the full user utterance action scheme of 61 tags, but improved when we compared only the top-level categories. Agreement in both cases was higher with the first author than between A2 and A3, which perhaps suggests confusion with different aspects of the scheme that need to be clarified in the manual and with further discussion. While these numbers could be improved by refining the definitions and guidelines, they are a very good starting point.

4.3. Analysis

The data resulting from this method of paired annotation allows us to look at the dialogue performance data from a number of different perspectives — some yielding results that are potentially comparable with other systems using the same method, others yielding results that are more useful for internal evaluation and developmental purposes. Each evaluation perspective is discussed below. An example user dialogue with Moleno is shown in figure 1.

4.3.1. Comparative Evaluation

In total, 100 dialogue interactions were annotated. The dialogues spanned 19 days of logs, with a total of 1,479 utterances, 785 (53.1%) by the agent. 335 (42.7%) of the agent's utterances were tagged as taking initiative. There were a total of 678 user utterances with evaluative ratings (the remainder were segments continuing a user utterance, which required no code). An overview of the results is shown in Table 5.

Rating	Result
3	167 (24.6%)
NR3	211 (31.1%)
2	67 (9.9%)
RR	73 (10.8%)
NR1	65 (9.6%)
1	95 (14%)
Total	678

Table 5: Overview of Response Ratings

Since the evaluation scheme has a variety of ratings, including silence as well as explicit responses, we create an analogy to precision and recall in order to make cross-evaluation comparisons easier to perform. "Appropriateness Rating" (AR) is analogous to recall, and expresses an overall sense of the fully appropriate reactions in the dialogue. It is calculated as AR = ('3' + 'NR3') / Total. "Response Precision" (RP), is analogous to precision, and gives a sense of the functional and conceptual precision of the agent's utterance responses. RP = "3"/ ("3"+ "2" + "RR" + "1"). The overall ratings for the 100 dialogues are as follows:

AR = 0.56
RP = 0.50

Considering only the quality of ratings when the agent did not respond, we may get a sense of whether his lack of responses were due to error or a balanced strategy in design. Here, AR^{NR}= NR3/(NR3 + NR1), and seems quite good, at 0.764. It is possible, however, that some features bias this dialogue domain toward silence being an unusually appropriate thing. So to test this, we considered the data from action categories where it seemed very desirable to respond, versus not.

Intuitively, it is generally bad to not respond to information requests, requests, statements about the domain, or statements initiating topics. These intuitions were confirmed by a lack of NR3 ratings for utterances in these categories. In contrast, categories where 'no response' is strongly desirable included utterances addressed to other (usually human) participants, exclamations, expressions of politeness, generic domain feedback and reactive statements. This was again confirmed, by a relative lack of NR1 ratings for these user utterances.

To compare the agent's performance in these contrasting situations, we calculated the number of cases where the agent gave an utterance response (of any quality) to a user's action vs. where he gave no response. The results in table 6 are not straight performance measures, but strongly confirm the agent's good performance on this aspect of dialogue.

Situation	Utterance	Silence	Total
Very Desirable to Respond	217 (87%)	32 (13%)	249
Very Desirable to be Silent	33 (27%)	190 (73%)	259

Table 6: Situational Performance to Respond or be Silent

In multiparty dialogues, addressee identification is an important feature of the agent's performance. Even though the agent did not have a specific component designed to identify addressees, we can evaluate its objective performance on utterances addressed to others in the data to determine whether the delay strategy implemented in multi-party dialogue situations is effective. There were a total of 109 utterances addressed to others in the dialogues, and the agent's AR = 0.734.

Speaker	Utterance	Code
Moleno	Good evening, I’m Lt Moleno, the Army Island Staff Duty Officer.	I
	Here is a calling card if you ever need to call me.	I
	How can I help you?	I
User	hold on	RMW-NR3
	(pause, 1min 26 seconds)	
User	hey i wanna go sky diving	SDP-3
Moleno	Jumpmaster Neiro at the Skydiving Center will help you.	
User	i wanna fly that black thing	SDP-RR
Moleno	Could you rephrase that?	
User	the airplane do you compute	SDP-1
Moleno	I don’t know, but I will find out and get back to you.	
Moleno	I will try to help in any way I can, but I like giving directions the most.	I
	I can give you a tour around the islands if you want.	I
User	ok where do i go to skydive	QDL-2
Moleno	Jumpmaster Neiro at the Skydiving Center will help you.	
User	take me there	RDX-3
Moleno	Just follow me.	

Figure 1: Example Coded Dialogue

The method as described so far should be useful in progress evaluations, where we compare different versions of the system. Also, as the base units for measuring the performance are unrelated to system-internal components, it could be used to compare how well different versions of system components perform on the same input data, without running further user testing. We leave further questions of comparative evaluation (application to other systems and comparison across domain types) to the discussion in section 5.

4.3.2. Developmental Evaluation

An advantage of the hierarchy — or multiple levels of granularity of action specification — in the descriptive coding method is that it allows flexibility for examining the performance data from different perspectives, by clustering ratings by different relations of categories, such as by general dialogue act type, or by domain-related utterances. Agents in publically available spaces increasingly need to deal with dialogue outside of their domain, regardless of their design goals. But it is good to be able to distinguish the system's performance, based on this dimension. In addition, even within the domain, there is often a difference between the domain characterization in agent design and the actual domain in use. Using the action categories of the user's utterances, we considered the agent's performance on these dimensions. 'Actual Domain' covers all user actions that were classified as in domain—this includes both domain specific actions, that is, functionally or topically, they addressed elements of Army Island and its activities in some manner, as well as some generic actions necessary to the domain communication that were included in design, such as greetings, closings, etc. 'Design Domain' includes only the subset of these actions that the agent specifically was designed to respond to. 'Domain Oversight' is the remainder— specifically, the domain actions not anticipated in the original design. The agent's performance in each of these cases is shown in Table 7.

Action Range	AR	RP
Actual Domain Performance	0.605	0.512
Design Domain Performance	0.654	0.571
Domain Oversight Performance	0.216	0.083

Table 7: Domain Related Agent Performance

There is a very large gap between the design performance and the oversight. These are likely an indication of the most pressing data or strategies that need to be considered in further system revisions. Looking objectively at performance, we can also determine how well the designers have covered the domain. The performance figures between 'design domain' and 'actual domain' are much closer largely because the domain data was fairly well covered. There were 461 utterances total within the 'actual domain.' Of these, 410 fell into 'design domain', and only 51 into 'oversight'.

While it is most important that Moleno perform well on actions relating to his design goals as greeter/ guide, the fact that he is situated in a public domain, and where users do not necessarily know he is an agent, ideally he should also perform reasonably well on any utterances put to him. The lack of knowledge about what users in a conversational domain are likely to say is a significant problem in the developmental process and early performance of more complex conversational agents (Bernsen & Dybkjaer, 2004; Robinson et al., 2008). Having a tagged corpus in mid-stage development can greatly help this process, and the action typology can allow developers to prioritize how, and in what order, to deal with this information, based on their working goals.

5. Discussion

While the utility of this method for developmental evaluation should be clear, the question remains how well it can be generalized to other systems, both for internal evaluation and comparative evaluation. The action description scheme would require some modification for each new domain — certainly in the lower level and domain descriptive categories, and likely some addition or deletion of other more generic dialogue acts as well. In principle, the evaluative scheme is directly applicable to any dialogue system, or even human dialogue performance. In practice, the resulting scores could only be compared directly to a similar dialogue situation, or at least a domain with a similar level of complexity, as domain complexity will clearly affect performance rates. While there are no well-defined measures of complexity in dialogue (another place where speech recognition is ahead of dialogue evaluation), the descriptive scheme here could possibly be used to help define relative complexity. Complexity in dialogue systems is typically discussed from the system perspective — a mixed initiative system has more complexity than a system relying solely on agent initiative. But this could equally be defined from the perspective of the range of acts the user performs in interacting with the system. With further annotation along these lines in different domains, we might anticipate that both the range of high level acts and the depth of variation within an action category would help to clarify some comparative scale of the complexity of the agent's task in dialogue performance.

6. Acknowledgements

We would like to thank members of our team who helped design and implement Staff Duty Officer Moleno, particularly Dusan Jan, Anton Leuski, Jackie Morie, as well as the creators of the Army Island in Second Life, and the users who interacted with Moleno in this early phase of development.

The project described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

7. References

Artstein, R., Gandhe, S., Gerten, J., Leuski, A., Traum, D. (2009). Semi-formal evaluation of conversational characters. In O.Grumberg, M. Kaminski, S. Katz, & S.Wintner (Eds.), *Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday, volume 5533 of Lecture Notes in Computer Science.* pages 22-35. Berlin: Springer, pp. 22-35.

Bernsen, N.O., Dybkjaer, L. (2004). Domain-Oriented Conversations with H.C. Andersen. In *Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems (ADS), Lecture Notes in Artificial Intelligence 3068.* Heidelberg: Springer Verlag, pp. 142-153.

Bernsen, N.O., Dybkjaer, L., Minker,W. (2007). Spoken dialogue systems evaluation. In L. Dybkjaer, H. Hemsen, & W. Minker (Eds.), *Evaluation of Text and Speech Systems.* Springer, pp. 187-219.

Bohlin, P., Bos, J., Larsson, S., Lewin, I., Matheson, C., Milward, D. (1999). Survey of Existing Interactive Systems. *Task-Oriented Instructional Dialogue, TRINDI Technical Report, LE4-8314.*

Danieli, M., Gerbino, E. (1995). Metrics for Evaluating Dialogue Strategies in a Spoken Language System. In *Working Notes AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pp. 34-39

Edlund, J.,Gustafson, J., Heldner, M., Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication*, 50, pp. 630-645.

Foster, M.E., Giuliani, M., Knoll, A. (2009). Comparing Objective and Subjective Measures of Usability in a Human- Robot Dialogue System. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP,* pp. 879–887

Holzapfel et al 2008. "Steps to Creating Metrics for Human-like Movements and Communication Skills (of Robots), in: Catherina R. Burghart & Aaron Steinfeld (Eds.), *Proc. of Metrics for Human-Robot Interaction, a Workshop at ACM/IEEE HRI 2008, Amsterdam, the Netherlands, 12 March 2008, Technical Report 471* Hatfield, UK: University of Hertfordshire

Jan, Dusan, Antonio Roque, Anton Leuski, Jackie Morie, & David R. Traum. (2009). A Virtual Tour Guide for Virtual Worlds. In *Proceedings of Intelligent Virtual Agents 2009*, pp. 372-378.

Kamm, C.A., Walker, M.A., Litman, D.J. (1999). Evaluating Spoken Language Systems. In *Proceedings of American Voice Input/Output Society(AVIOS), May 1999*, pp. 187-197.

King, M. (2007). General principles of user-oriented evaluation. In L. Dybkjaer, H. Hemsen, & W. Minker (Eds.), *Evaluation of Text and Speech Systems.* Springer, pp. 125-161.

Leuski, A., Traum, D., (2010). NPCEditor: A Tool for Building Question-Answering Characters. In *Proceedings of LREC 2010 — 7th Language Resources and Evaluation Conference (Malta, May 2010).*

Möller and Ward, 2008, A framework for model-based evaluation of spoken dialog systems

Paek, T. (2007). Toward Evaluation that Leads to Best Practices: Reconciling Dialogue Evaluation in Research and Industry. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*. Rochester, N.Y.: Association for Computational Linguistics, pp. 40-47.

Robinson, S., Traum, D., Ittycheriah, M., Henderer, J. (2008). What would you ask a conversational agent? Observations of Human-Agent dialogues in a museum setting. In *Proceedings of LREC 2008- 6th Language Resources and Evaluation Conference (Marrakech, Morocco, May 2008)*

Traum. D. (2000). 20 questions for dialogue act taxonomies. *Journal of Semantics*, 17(1), pp. 7-30.

Traum, D., Robinson, S., Stephan, J. (2004). Evaluation of Multi-Party Virtual Reality Dialogue Interaction. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC),* pp. 1699-1702.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, 59, pp. 433-460.

Walker, M., Litman, D., Kamm, C., Abella,A. (1997). Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings ACL-97.*

Walker, M., Passonneau, R. (2001) DATE: A dialogue act tagging scheme for evaluation of spoken dialogue systems. In *Proceedings of the First International Conference on Human Language Technology Research,* , Mar 18-21, 2001, San Diego, pp 1-8.