Dialog Management for Rapid-Prototyping of Speech-Based Training Agents

Victor C. Hung, Avelino J. Gonzalez, Ronald F. DeMara
University of Central Florida
Orlando, Florida
victor@isl.ucf.edu, gonzalez@ucf.edu, demara@mail.ucf.edu

ABSTRACT

Speech-based training agents can be described as virtual humans posing as interactive training characters with the capability to communicate in a spoken conversational manner. While creating this technology, developers face two stumbling blocks: 1) modeling the agent and its training knowledge is a time-consuming and tedious task, and 2) modern speech recognition software suffers from high Word-Error Rates caused by noisy environmental conditions. This paper presents a dialog management architecture that addresses these problems using the Context-Based Reasoning paradigm. The system minimizes the time necessary to build the training knowledge in the instructional agent, as well as tolerates the relatively high Word-Error Rates related to automatic speech recognition. Ultimately, these advantages lead to quick development of speech-based training agents. The dialog manager was directly implemented into the LifeLike Avatar, an embodied conversational agent funded by the National Science Foundation. A set of quantitative results is presented to reflect the effectiveness of the system.

ABOUT THE AUTHORS

Victor C. Hung is a Ph.D. student in Computer Engineering at the University of Central Florida (UCF). His work entails developing a natural language dialog system based on the Context-Based Reasoning paradigm.

Avelino J. Gonzalez is a Professor of the School of Electrical Engineering and Computer Science at UCF. His research interests focus on the areas of artificial intelligence, context-based behavior and representation, temporal reasoning, intelligent diagnostics and expert systems.

Ronald F. DeMara is a Professor in the School of Electrical Engineering and Computer Science at UCF. His research interests are in Computer Architecture with emphasis on Real-time Intelligent Systems.

Dialog Management for Rapid-Prototyping of Speech-Based Training Agents

Victor Hung, Avelino Gonzalez, Ronald DeMara
University of Central Florida
Orlando, Florida
victor@isl.ucf.edu, gonzalez@mail.ucf.edu, demara@mail.ucf.edu

INTRODUCTION

With the appearance of ELIZA (Weizenbaum, 1966), the technology world was introduced to conversation-based interaction with a computer. Additionally, advances in automatic speech recognition (ASR) technology have resulted in hopes that voice interaction will someday replace the keyboard and mouse with a microphone. The melding of conversation agents with ASR has made speech-based agents almost a reality. In turn, these entities are very likely to find their way into the training realm.

Nevertheless, conversational, speech-based interaction in any type of embodied agent remains a difficult problem to solve. Dialog systems that catch the spoken words, make sense of them, and compose an intelligent and natural response must be built robustly and reliably. In fact, Kang et al (2009) assert that a challenging aspect of dialog systems is processing speech-based input. Environmental factors, such as interfering background noise, can muddle a machine's ability to pick up spoken utterances. Additionally, a non-native speaker could give a grammatically inaccurate response, causing even more confusion when an agent must determine the user's needs. Speech recognizer systems have seen Word-Error Rates (WER) peak out at 30%, under controlled conditions. (Kang et al, 2009) This is an unacceptable level of integrity if a developer wishes to use full syntactic processing of user inputs.

This paper presents a method to elevate the level of speech-based discourse to a new echelon of naturalness in embodied conversation agents (ECA) by exhibiting a tolerance to high WERs in speech understanding systems, yet providing answers to asked questions, without pre-linking questions to answers. An additional aspect of producing embodied training agents is the amount time required to develop the knowledge base used to generate the appropriate response. There is a need for a generalized ECA knowledge infrastructure, yielding a quicker domain-independent agent development lifecycle.

Hence, this general problem yields two specific problems to be addressed: 1) tolerating the limitations of current untrained ASR technology, and 2) developing a domain-independent knowledge management system. In this paper, we show that a domain-independent, context-driven conversational discourse infrastructure applied to speech-based conversational dialog systems provides an effective level of natural language understanding for an assistive interaction between an ECA and a human in spite of low performing ASR systems. The intent of this work is to pass on the benefits of such a dialog manager to the realm of embodied training agent design.

BACKGROUND

Two major themes permeate the work presented in this paper: conversation agent design and context-based methods. This section will present a brief discussion of each of these topics.

Conversation Agent Design

Conversation agents, or chatbots, represent a specialized field of natural language processing (NLP) whose purpose is to provide a method of interaction that resembles a conversational exchange between two humans. Additionally, *assistive* dialog systems exist to serve a particular purpose, such as providing information to its users or aiding them with completing a certain task. The channel of communication between the user and the dialog system may be based on text, on speech and gestures, or on a combination of all three.

Conversation agents began life with ELIZA (Weizenbaum, 1966), followed by the development of PARRY (Colby, 1973) and SHRDLU (Winograd, 1980). Chatbot development experienced a period of inactivity until the 1990's with the introduction of ECAs, where they were paired with a physical embodiment. (Gorin et al, 1997) (Casell et al, 2000) (McBreen and Jack, 2000) (Massaro et al, 2001)

(Catrambone, 2002) (Béchet et al, 2004) Innovations in ECA technology would follow, such as virtual storytellers (Tarau and Figa, 2004), museum guides (Kopp et al, 2005), emotional entities (Alm et al, 2005), caring agents (Bickmore and Picard, 2004) (Turunen et al, 2008) (Bickmore et al, 2009) (Ferguson et al, 2009) (Galescu et al, 2009) (Kenny et al, 2009), and fully-embodied realism (Kenny et al, 2007). Training ECAs have even emerged very recently, as seen in Hassan. (Gandhe et al, 2009)

Context-Based Methods

Context-based methods refer to the techniques used by a machine to drive behavior based on the immediate environmental state, or current context. Context-based reasoning (CxBR) formalizes this concept as a paradigm of agent behavior in which only a subset of an entity's total knowledge is active at any one time (Gonzalez and Ahlers, 1998) (Gonzalez et al, 2008), reflective of how humans themselves only require a fraction of their knowledge for any given situation.

In general, NLP problems can easily be enhanced through contextually-driven methods (Porzel and Strube, 2002), such as those found in spoken language translation (Levin et al, 1995), semantic clarification (Kladke, 1989) (Towhidnejad, 1990) and knowledge modeling (Porzel et al, 2006). Perhaps the group of natural language protocol researchers that has benefited the most from contextualization is the ASR community (Young, 1989) (Serridge, 1997) (King et al, 1998) (Goulian et al, 2003) (Fügen et al, 2004) (Yan and Zheng, 2004) (Sarma and Palmer, 2004) (Lieberman et al, 2005) Sammut's ProBot (2001) touches upon using context in conversation agent discourse, which uses contexts when unexpected utterances are received.

This section brings to attention the general body of chatbot work that has surfaced since its infancy. It also mentions the extent to which context-based methods have been used for the sake of natural language systems. However, there lacks a presence of combining context-based methods with dialog systems. The premise of this work is to provide a linkage between these technologies, producing a conversation agent whose dialog management is driven by context-based methods.

CONTEXT-BASED DIALOG SYSTEM

Devising a new dialog system involves three major design decisions: user input processing method, knowledge management, and agent response discourse mechanism. The aggregation of all three sub-system selections results in the final approach of the prototype, the CONtext-driven Corpus-based Utterance Robustness dialog manager (CONCUR) prototype. CONCUR was incorporated into the LifeLike Avatar (DeMara et al, 2008) ECA. The next sections describe the inner-workings of the CONCUR prototype.

Input Processor

The first component of CONCUR is the Input Processor. This module parses the raw user utterance for contextual keyphrases. The result from the speech recognizer is chunked into phrases, which begins with a word-for-word Parts-of-Speech tagging. This procedure is performed using an NLP toolkit. The utterance phrase chunks are then filtered for noun and verb phrases, discarding the remainder of the sentence.

The Input Processor also contributes to annotating the Knowledge Manager's corpora with keyphrase tags. As part of the corpus pre-processing routine, an NLP treatment of corpus data is performed, providing an automatically generated keyphrase list for each contextual layer. The next section delves further into the Knowledge Manager.

Knowledge Manager

A major feature of CONCUR remains its dependency on contextual relevance. This concept speaks to the idea that two data points may be within contextual proximity of each other if they share some form of conceptual commonality. Hence, contextualization requires a pre-defined set of related data. For dialog-based systems, contextualization exists when groups of words maintain conceptual relationships with each other. These lingual relationships are contained in the Knowledge Manager portion of the CONCUR architecture. Knowledge bases used by CONCUR all reflect a pre-established contextual relationship mapping. This is done by using a contextual layering system of organizing information, a format similar to that of an outline or an encyclopedia entry. Hence, all responses that will be used as spoken replies by the conversation agent are pre-annotated in the knowledge base with a contextually-driven tagging system derived from its outline depth.

The Knowledge Manager consists of three sources of data: user data, conversational knowledge, and domain-specific knowledge. The domain-specific knowledge features the expertise of the National Science

Foundation (NSF) Industry/University Cooperative Research Center program (I/UCRC). The scope and depth of the domain-specific knowledge is modeled after that of a traditional expert system. (Gonzalez and Dankel, 1993) The entirety of the conversational knowledge base encompasses all of the agent quips that do not reflect any expertise, but rather serve as transitioning actions. The user profile database serves as a repository of individualized account data, and it reflects the importance of memory during a conversation. (MacWhinney et al, 1982)

The strength of the Knowledge Manager is its ability to fetch contextualized knowledge. Contextualized knowledge refers to a cross-section of any of the three knowledge sources that is relevant for the active context of the conversation. Once the dialog system determines the context of the conversation, knowledge that is labeled with the current context is elicited as valid information for the conversation and funneled into the contextualized knowledge base. Once this subset of information is established, the dialog manager can then work with a manageable portion of the entire knowledge base.

Discourse Model

CONCUR uses a conversation discourse model driven by the CxBR paradigm. With knowledge of the current state of the conversation, the discourse model pieces together the information of the Input Processor and the Knowledge Manager, as well as adds its own CxBR devices to provide an appropriate reply to the user.

The underlying theme of the discourse model is the supervision of conversation goals. In essence, a spoken dialog between parties is a sequence of passing goal-oriented statements to one another. (Grice, 1975) The intent is to achieve some form of resolve for each of these exchanges, otherwise viewed as completing goals. For this paper, the dialog manager is charged with managing these goal completion tasks. The detection of goals is performed in the inference engine. Contexts directly correlate to reaching specific goals. Servicing of goals is the work of traversing the context topology until the mission is completed. This goal management comprises the processes that recognize and satisfy the interlocutor's needs as conveyed by her/his utterances. CONCUR's goal management involves two parts: 1) goal bookkeeping and 2) context topology.

The Goal Bookkeeper maintains the goal-based activities of conversation agent, and it consists of two parts: the Inference Engine and the Goal Stack. The primary purpose of the Inference Engine is to recognize the user's immediate goals. Goal recognition refers to the process of analyzing user input utterances to determine the proper conversational goal that is to be addressed. This is analogous to the context activation process in CxBR methods. The CONCUR Inference Engine takes the user utterance and performs a keyphrase matching among the current set of relevant contexts. This means that a strong knowledge base must be in place for proper goal recognition. It is up to the Inference Engine to determine if the user has remained within the current conversational context, or if there has been a topic switch.

The Goal Stack directly manages the conversation flow. This mechanism serves as an agent's short-term memory during a conversation. Its job is to ensure that all contexts that are introduced into a dialog exchange are attended to and eventually brought to closure by the end of the conversation. The Goal Stack performs its context management immediately upon the Inference Engine's selection of the current context.

CONCUR's Context Topology gives structure to its entire set of speech. This includes the transitional actions when moving between contexts when a goal shift is detected. The Context Topology organizes the contexts that represent the set of behaviors through which the system will respond. Two major types of contexts make up these mixed-initiative dialog-based actions: Agent Goal-Driven Contexts, and User Goal-Driven Contexts. These two sides of conversational contexts reflect Grice's (1975) treatment on goals in dialog. Agent Goal-Driven Contexts pertain to those actions needed for the avatar itself to perform its duties as an interfacing agent. User Goal-Driven Contexts refer to those behaviors needed to help support the user fulfill her/his needs.

This section gave a description of the CONCUR prototype's basic architecture. The next section describes how it will be test for effectiveness as a framework for building speech-based training agents.

EVALUATION AND RESULTS

Evaluation of chatbots has always remained a controversial topic, as it is unclear on how to quantitatively describe how well a conversation agent performs or how naturally it responds. Current research has found success in using both quantitative and quality metrics to make relative comparisons between

conversation agents. (Semeraro et al, 2003) (Rzepka et al, 2005) (Shawar and Atwell, 2007) An assistive dialog system proves its effectiveness under the light of two primary objectives: 1) dialog performance, and 2) task success. (Walker et al, 1997) (Dybkjær and Bernsen, 2001) Walker et al (1997) further break down the dialog performance to efficiency measures and quality measures. Efficiency costs refer to the quantitatively-measurable resource consumption needed to accomplish a single task or sub-task. Quality costs measure the actual conversational content. These metrics may be recorded quantitatively or qualitatively. Table 1 delineates the relevant cost metrics for CONCUR, as inspired by Walker et al (1997), Stibler and Denny (2001), Charfuelán et al (2002), and Hassel and Hagen (2005).

Table 1. Dialog cost metrics

Metric	Type	Data Collection Method
Total elapsed time	Efficiency	Quantitative Analysis
Total number of user turns	Efficiency	Quantitative Analysis
Total number of system turns	Efficiency	Quantitative Analysis
Total elapsed time per turn	Efficiency	Quantitative Analysis
Word-Error Rate	Quality	Quantitative Analysis
Total number of out-of-corpus misunderstandings	Quality	Quantitative Analysis
Total number of general misunderstandings	Quality	Quantitative Analysis
Total number of inappropriate responses	Quality	Quantitative Analysis
Total number of user goals	Quality	Quantitative Analysis
Total number of user goals fulfilled	Quality	Quantitative Analysis
Conceptual accuracy	Quality	Quantitative Analysis
Conversational accuracy	Quality	Quantitative Analysis
Usefulness	Quality	Questionnaire
Naturalness	Quality	Questionnaire

Efficiency Metrics pertain to those interaction traits that can be empirically observed, with no need for quality interjection. A second set of results, called the Quality Metrics, was observed using both Quantitative Analysis and Questionnaire-based data.

The Quality metrics were collected after a user interaction was completed, where the transcript of the experiment was analyzed. At the conclusion of each experimental interaction, the users were given an exit survey. On this instrument, each question is a statement in which the user provides her/his level of agreement, where a '1' rating is 'I disagree' and a '7' corresponds to 'I agree.' The "Naturalness" statements aim to garner whether the user was able to experience a natural conversational exchange, while the "Usefulness" statements check if the agent was able to perform as a capable information deployment tool.

The Naturalness statements are as follows:

- If I told someone the character in this tool was real they would believe me.
- The character on the screen seemed smart.
- I felt like I was having a conversation with a real person.
- This did not feel like a real interaction with another person.

The Usefulness statements consist of:

- I would be more productive if I had this system in my place of work.
- The tool provided me with the information I was looking for.
- I found this to be a useful way to get information.
- This tool made it harder to get information than talking to a person or using a website.
- This does not seem like a reliable way to retrieve information from a database.

Experimental Design

During the experimentation process, a single conversational scenario was employed on four different assistive dialog systems. The idea was to provide a comparison study between different conversation agents, with special attention to two of CONCUR's traits: 1) ability to provide usefulness as a speech-based ECA, and 2) capability to maintain effectiveness over different expert domains.

The first experiment involved the speech-based CONCUR dialog manager. A photorealistic animated embodiment, the LifeLike Avatar (DeMara et al, 2008), was used as the agent interface and it was fully operational with CONCUR as its dialog manager. This ECA represented the performance of speech-based systems specializing in context-sensitive dialog management, which is the primary intent of this paper. A rich corpus pertaining to an NSF research funding effort known as the Industry/University Collaborative Research Centers (I/UCRC) program, corpus was loaded into this agent. The speech-based component of the CONCUR reflects a real-world spoken conversation situation with a less-than-optimal WER. Thirty trials were conducted in this experiment using 21 male 9 female participants. Six of these trials involved a non-

native English speaker, and approximately half of the test subjects were already familiar with the NSF I/UCRC program.

The second experiment engaged the user with another speech-based agent, but one whose speech action engine does use an open dialog discourse, but rather, more of a menu-driven model. The AlexDSS Expert System (Sherwell et al, 2005) knowledge engine was directly implemented as the dialog engine for the LifeLike Avatar. This ECA's method of dialog management resembles that of an automated phone, using a highly constrained style of user input expectation. Thirty trials were performed for this experiment, resulting in 30 collected exit surveys and 20 speech transcripts for analysis. Of these transcripts, 14 belonged to male participants, and 6 came from female users. Five of the recorded experiments were completed by non-native English speakers and 40% of the analyzed data were from personnel familiar with the NSF I/UCRC program.

In the third experiment, current events knowledge was fitted into the CONCUR framework. This experiment evaluates the domain-independence aspect of CONCUR. The corpus was built from a selection of current news articles from the World Wide Web pertaining to U.S. and international news stories, sporting events, and personal health issues. To eliminate ASR errors, a purely text-based user input system was implemented. This also meant that no physical embodiment of the agent was used, such as that used in the LifeLike Avatar. Twenty surveys and chat transcripts were collected from this experiment, consisting of 14 male and 6 female participants.

The final experiment used a text-based CONCUR dialog manager to simulate ideal ASR conditions. User inputs taken from Experiment 1's transcripts were entered into a text-based version of CONCUR. The responses from the agent were recorded to reflect a version of CONCUR that does not have input errors associated with ASR facilities. This experiment gave a baseline for measuring CONCUR's effectiveness. Since no extra participants were needed for this experiment, no surveys data was necessary.

Table 2 gives a summary of all of the experiments featured in this paper, establishing the differences in dialog manager selection, agent style, domain expertise, and number of trials conducted.

Table 2. Experiment design

Experiment	Dialog Manager	Agent Style	Domain	Surveys/ Transcripts Collected
1	CONCUR	LifeLike Avatar	NSF I/UCRC	30/30
2	AlexDSS	LifeLike Avatar	NSF I/UCRC	30/20
3	CONCUR	Chatbot	Current Events	20/20
4	CONCUR	Chatbot	NSF I/UCRC	0/20

Results

The four experiments were conducted and their results are presented in this section. Table 3 shows the aggregate survey results for the first three experiments, those that that utilized the survey instrument. This table only depicts the results for the average of the Naturalness and Usefulness statements. From these results, it is observed that the NSF I/UCRC CONCUR LifeLike Avatar scored the highest in both categories.

Table 3. Survey results

Experiment	Naturalness	Usefulness
1	4.14	4.51
2	4.02	4.47
3	2.40	3.38

Table 4 shows the efficiency metrics collected from the four experiments. These metrics deal with the measurable, non-quality judgments recorded by each agent. The WER metric reports how well the ASR performed. In this table, it is exhibited that both speech-based ECAs yielded less than 45% accuracy in detecting user speech, a number much less than those found in recent ASR efforts. (Kang et al, 2009)

Table 4. Efficiency metrics results

	Experiment			
	1	2	3	4
Total Elapsed Time (min)	3:20	3:36	4:03	2:52
Number of User Turns	10.90	13.35	8.85	10.10
Number of System Turns	11.90	14.35	9.85	11.10
Elapsed Time Per Turn (s)	6.10	4.15	9.37	6.11
WER	58.48%	60.85%	0.00%	0.00%

Table 5 gives the results of the quality metrics. In these metrics, each chat transcript was analyzed for misunderstanding, erroneous agent responses, and context goal satisfaction. The final two columns, Goal Completion Accuracy and Conversational Accuracy, give a quantitative account of each agent's usefulness and naturalness, respectively. Goal Completion Accuracy gives the percentage of user information requests that were adequately fulfilled by the agent. Conversational Accuracy accounts for the percentage of system responses that are not classified as a general misunderstanding or an erroneous reply.

From this table, the CONCUR chatbot with the current events corpus achieved the highest Conversational Accuracy, despite exhibiting the lowest Goal Completion Accuracy. The LifeLike Avatar CONCUR with the NSF I/UCRC corpus produced the worst Conversational Accuracy, and its Goal Completion ability were nearly as accurate as that of the AlexDSS NSF I/UCRC ECA.

Table 5. Quantitative analysis of quality metrics

	Experiment			
	1	**2**	**3**	**4**
Out-Of-Corpus Misunderstanding Rate	6.15%	0.29%	17.45%	6.77%
General Misunderstanding Rate	14.49%	9.51%	0.00%	7.48%
Misunderstanding Rate	20.64%	9.80%	17.45%	14.25%
Error Rate	21.81%	8.71%	16.46%	16.68%
Goal Completion Accuracy	60.48%	63.29%	48.08%	68.48%
Conversational Accuracy	63.93%	81.78%	83.54%	75.34%

Discussion of Results

The impetus of the experiments in this paper was to weave a story about building a dialog manager that could overcome ASR limitations and provide domain-independent knowledge management. This section discusses the experimental results with these two themes in mind.

Speech Recognition Limitations

To assess CONCUR's ability to handle ASR limitations, the results from Experiment 1 were compared against those of the other experimental agents. Table 6 gives a comparative look at a subset of metrics from the experimentation.

Table 6. NSF corpus agent comparison

		Experiment		
		1	**2**	**4**
Survey Results	**Naturalness**	4.14	4.02	n/a
	Usefulness	4.51	4.47	n/a
Quant. Metrics	**WER**	58.48%	60.85%	0.00%
Quant. Analysis	**Goal Completion Accuracy**	60.48%	63.29%	68.48%
	Conversational Accuracy	63.93%	81.78%	75.34%

In this table, it is observed that the avatar-based NSF I/UCRC CONCUR from Experiment 1 yielded a similar WER and a much lower Conversational Accuracy when compared to the AlexDSS avatar from Experiment 2, yet it still scored higher in the user survey ratings for Naturalness and Usefulness. Experiment 4's text-based NSF I/UCRC CONCUR served as the baseline for a perfect speech recognition system. Even with these ideal ASR conditions, the speech-based CONCUR was still able to achieve Goal Completion and Conversational Accuracy numbers within those of its text-based counterpart. CONCUR is based on an open dialog method, while AlexDSS aligns to a more direct, automated phone operator style. In relation to the results, this difference in user response expectation would account for the worsened ASR results for CONCUR, as well as its deteriorated Conversational Accuracy.

To put CONCUR's performance in perspective, four recent projects were selected for comparison, as depicted in Table 7. In this table, CONCUR's user rating statistics in Naturalness and Usefulness edge out those of the speech-based ECAs Amani and Hassan in a similar survey-based assessment measure. (Gandhe et al, 2009) Additionally, the Goal Completion Accuracy of CONCUR is comparable to that of the spoken Digital Kyoto agent (Misu and Kawahara, 2007), despite the presence of twice as many word errors. Finally, the text-based TARA (Schumaker et al, 2007) yielded a poorer average Conversational Accuracy rating when compared to the speech-based CONCUR agent.

Table 7. CONCUR versus other agents

Agent	Natural	Useful	Avg. WER	Goal Comp. Acc.	Conv. Acc.
Speech-based NSF I/UCRC CONCUR	4.14	4.51	58.48%	60.48%	63.93%
Amani (Gandhe et al, 2009)	3.09	3.24			
Hassan (Gandhe et al, 2009)	3.55	4.00			
Digital Kyoto (Misu and Kawahara, 2007)			29.40%	61.40%	
TARA (Schumaker et al, 2007)			0.00%		54.00%

From the above comparisons of CONCUR with the AlexDSS ECA, as well as with other conversation agent research efforts, it was determined that a comparable level of usefulness could be achieved using a context-driven discourse method. Furthermore, the subjective assessments regarding the naturalness dictated that the ECA-based CONCUR's ability to conduct a conversation was slightly more natural than some contemporary speech-based agents. Hence, from these experiments, it was observed that CONCUR could perform adequately in the presence of ASR limitations.

Domain-Independence

The results of the CONCUR agents from Experiment 1 and Experiment 3, as seen in Table 8, pertain to CONCUR's domain-independence. It is quite noticeable that a decrease in performance was experienced when moving from the speech-based interface to a text-based one, as well as the shift from the NSF I/UCRC expertise to the Current Events knowledge. Perhaps most concerning are the drops in Naturalness User Rating and Goal Completion Accuracy. Nevertheless, the current events CONCUR chatbot in Experiment 3 agent was able to improve upon the Conversational Accuracy metric.

There is, however, a problem with comparing the performance metrics of these two agents against each other, as they both could garner better gains due to the nature of their input methods or their corpus differences. For example, general misunderstandings in text-based agents are eliminated because any errors associated with misconstrued ASR results do not come into play. On other hand, the speech-based agent may get better Naturalness ratings only because it resides within a full-blown ECA setup, such as the LifeLike Avatar, instead of a chat window.

Additionally, the wide range of topics in the Experiment 3 corpus seemed to cause a false assumption of agent omniscience in its users, a phenomenon not prevalent in the NSF I/UCRC-based Experiment 1. In both experiments, users were encouraged to engage in topic-based information requests. Experiment 1's tightly-knit expertise on a niche subject caused users to limit their information requests to I/UCRC-centric inquiries. The wider-scoped current events corpus of Experiment 3 made its users assume that the agent had more information than it was actually equipped to handle. This misunderstanding of the agent's expertise is presumed to be the primary cause of the decrease in Goal Completion Accuracy in the Experiment 3 chatbot.

The idea to take away from Table 8 is the fact that the same CONCUR agent infrastructure can still provide a usable and functionally acceptable dialog management experience regardless of any changes to its input method or corpus data, serving as the experimental basis for CONCUR's ability to support domain-independence.

Table 8. Comparison of CONCUR agents with different domains

Metric	Experiment 1	Experiment 3
Agent Type	ECA	Chatbot
Corpus Data	NSF I/UCRC	Current Events
Naturalness User Rating	4.14	2.40
Usefulness User Rating	4.51	3.38
Total Elapsed Time	3:20 min	4:03 min
Out-Of-Corpus Misunderstanding Rate	6.15%	17.45%
General Misunderstanding Rate	14.49%	0.00%
Misunderstanding Rate	20.64%	17.45%
Error Rate	21.81%	16.46%
Goal Completion Accuracy	60.48%	48.08%
Conversational Accuracy	63.93%	83.54%

Another aspect of domain-independence is the quick turnover time for creating a new agent knowledge base compares the turnover time for creating new domain expertise for a CONCUR agent versus the development time for other dialog systems.

From this table, it is easy to see that CONCUR's domain-independent knowledge management emphasizes its advantage as a rapid-prototyping tool for

ECA dialog creation for the training arena.

Table 9. Approximate knowledge base development turnover for dialog systems

Dialog System	Knowledge Management Method	Turnover Time
CONCUR	Corpus-based	3 Days
Marve (Babu et al, 2006)	Wizard-of-Oz Knowledge Base	18 Days
Amani (Gandhe et al, 2009)	Question-Answer Pairs Rule Base	Weeks
Sergeant Blackwell (Robinson et al, 2008)	Wizard-of-Oz Knowledge Base	7 Months
Sergeant Star (Artstein et al, 2009)	Question-Answer Pairs Rule Base	1 Year
HMIHY (Béchet et al, 2004)	Hand-modeled Knowledge Agent	2 Years
Hassan (Gandhe et al, 2009)	Question-Answer Pairs Rule Base	Years

In this section, the CONCUR infrastructure demonstrated its ability to effectively maintain its primary functionality as an assistive conversation agent regardless of its domain expertise. Additionally, a comparison of the knowledge base development times for different dialog systems saw that CONCUR's corpus-based knowledge management yielded the quickest turnover time.

SUMMARY

The research described here dealt with spoken interaction with a computer with emphasis on natural conversation flow for use in embodied training agents. Specifically, it presented a context-driven method of dialog management to fortify the robustness of assistive speech-based ECAs.

The particular areas of improvement were concentrated in two themes: overcoming ASR limitations and providing a domain-independent knowledge management system. An approach to building the context-driven dialog manager was presented with special emphasis on three primary design decisions: the input processing method, the knowledge management process, and discourse model.

A prototype of this approach was reflected in the CONCUR dialog system, whose architecture focused primarily on the use of contextual information to drive a conversation. Experiments for CONCUR were conducted to validate the two themes of context-based dialog management: ASR limitations and domain-independent knowledge management. The results consisted of quantitative metrics, survey responses, and quantitative analyses of quality data. Analyzing these data led to the experimental verification of the aforementioned themes.

The results from this work show that speech-based training agents can be effectively developed by driving the focus of conversation toward context-level processing instead of relying heavily on syntactic interpretations.

ACKNOWLEDGEMENTS

This research is supported by NSF Collaborative Research award 0703927.

REFERENCES

Alm, C., Roth, D., and Sproat, R. (2005). Emotions from Text: Machine Learning for Text-based Emotion Prediction. *HLT/EMNLP 2005*.

Artstein, R., Gandhe, S., Gerten, J., Leuski, A., and Traum, D. (2009). Semi-formal Evaluation of Conversational Characters. Languages: *From Formal to Natural: Essays Dedicated to Nissim Francez on the Occaision of his 65th Birthday*, pp. 22-35.

Babu, S., Schmugge, S., Barnes, T., and Hodges, L. (2006). What Would You Like to Talk About? An Evaluation of Social Conversations with a Virtual Receptionist. *6th International Conference on Intelligent Virtual Agents*.

Béchet, F., Gorin, A., Wright, J., and Hakkani-Tür, D. (2004). Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How May I Help You? *Speech Communication*, 42(2), pp. 207-225.

Bickmore, T. W., and Picard, R. W. (2004). Towards Caring Machines. *Computer Human Interaction*.

Bickmore, T., Pfeifer, L., and Jack, B. (2009). Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. *CHI 2009*, pp. 1265-1274.

Cassell, J., Ananny, M., Basu, A., Bickmore, T., Chong, P., Mellis, D., Ryokai, K., Smith, J., Vilhjálmsson, H., and Yan, H. (2000). Shared reality: Physical collaboration with a virtual peer. *Proceedings of CHI 2000*.

Catrambone, R. (2002). Anthropomorphic Agents as a User Interface Paradigm: Experimental Findings and a Framework for Research. *24th Annual Conference of the Cognitive Science Society*, pp. 166-171.

Charfuelán, M., Gómez, L. H., López, C. E., and Hemsen, H. (2002). A XML-based tool for evaluation of SLDS. *Proceedings of the Third*

International Conference on Language Resources and Evaluation.

Colby, K. (1973). Simulation of belief systems. *Computer Models of Thought and Language*.

DeMara, R., Gonzalez, A., Jones, S., Johnson, A., Hung, V., Leon-Barth, C., Dookhoo, R., Leigh, J., Renambot, L., Lee, S., and Carlson, G. (2008). Toward Interactive Training with an Avatar-based Human-Computer Interface. *Interservice/Industry Training, Simulation & Education Conference*.

Dybkjær, L., and Bernsen, N. (2001). Usability evaluation in spoken language dialogue systems. *Annual Meeting of the ACL Worshop on Evaluation for Language and Dialogue Systems*, 9.

Ferguson, G., Allen, J., Galescu, L., Quinn, J., and Swift, M. (2009). CARDIAC: An Intelligent Conversational Assistant for Chronic Heart Failure Patient Heath Monitoring. *AAAI Fall Symposium on Virtual Healthcare Interaction*.

Fügen, C., Holzapfel, H, and Waibel, A. (2004). Tight coupling of speech recognition and dialog management - dialog-context dependent grammar weighting for speech recognition. *INTERSPEECH-2004*, 169-172.

Galescu, L., Allen, J., Ferguson, G., Quinn, J., and Swift, M. (2009). Speech Recognition in a Dialog System for Patient Health Monitoring. IEEE International Conference on Bioinformatics and Biomedicine *Workshop on NLP Approaches for Unmet Information Needs in Health Care*.

Gandhe, S., Whitman, N., Traum, D., and Artstein, R. (2009). An Integrated Authoring Toolkit for Tactical Questioning Dialogue Systems. *IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

Gonzalez, A. J., and Dankel, D. D. (1993). *The Engineering of Knowledge-Based Systems Theory and Practice*. Prentice-Hall, Incorporated, Englewood Cliffs, New Jersey.

Gonzalez, A., and Ahlers, R. (1998). Context-based representation of intelligent behavior in training simulations. *Transactions of the Society for Computer Simulation International*, 15(4), pp. 153-166.

Gonzalez, A., Stensrud, B., and Barrett, G. (2008). Formalizing context-based reasoning: A modeling paradigm for representing tactical human behavior. *International Journal of Intelligent Systems*, 23(7), pp. 822-847.

Gorin, A. L., Riccardi, G., Wright, J. H. (1997). How may I help you. *Speech Communication*, 23, pp. 113-127.

Grice, H. P. (1975). Logic and conversation. In p. Cole and J. Morgan, editors, *Syntax and Semantics: Vol. 3, Speech Act*, pp. 43-58, Academic Press, New York.

Hassel, L., and Hagen, E. (2005). Evaluation of a Dialogue System in an Automotive Environment. *6th SIGdial Workshop on Discourse and Dialogue*.

Kang, S., Lee, S., and Seo, J. (2009). Dialogue Strategies to Overcome Speech Recognition Errors in Form-Filling Dialogue. *22nd International Conference on Computer Processing of Oriental Languages*, pp. 282-289.

Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsela, S., and Piepol, D. (2007). Building Interactive Virtual Humans for Training Environments. *I/ITSEC'07*.

Kenny, P., Parsons, T., Rizzo, A. (2009). Human Computer Interaction in Virtual Standardized Patient Systems. *HCI* (4), pp. 514-523.

King, S., Stephenson, T., Isard, S., Taylor, P., and Strachan, A. (1998). Speech recognition via phonetically featured syllables. *Proceedings of ICSLP'98*.

Kladke, R. R. (1989). A Mega-Heuristic Approach to the Problem of Component Identification in Automated Knowledge Generation. *M. S. Thesis, University of Central Florida*.

Kopp, S., Gesellensetter, L., Krämer, N.C., and Wachsmuth, I. (2005). A conversational agent as museum guide - design and evaluation of a real-world application. *Intelligent Virtual Agents*, 3661, pp. 329-343.

Levin, L., Glickman, O., Qu, Y., Gates, D., Lavie, A., Rose, C. P., Van Ess-Dykema, C., and Waibel, A. (1995). Using Context in Machine Translation of Spoken Language. *Proceedings of Theoretical and Methodological Issues in Machine Translation (TMI-95)*.

Lieberman, H., Faaborg, A., Daher, W., and Espinosa, J. (2005). How to Wreck a Nice Beach You Sing Calm Incense. *International Conference on Intelligent User Interfaces*.

MacWhinney, B., Keenan, J. M., and Reinke, P. (1982). The role of arousal in memory for conversation. *Mem Cognit*, 10(4), pp. 308-317.

Massaro, D., Cohen, M., Beskow, J., and Cole, R. (2001). Developing and evaluating conversational agents. *Embodied conversational agents*, pp. 287-318.

McBreen, H., and Jack, M. (2000). Empirical Evaluation of Animated Agents In a Multi-Modal Retail Application. *AAAI Fall Symposium: Socially Intelligent Agents - The Human in the Loop*, pp. 122-126.

Misu, T., and Kawahara, T. (2007). An Interactive Framework for Document Retrieval and Presentation with Question-Answering Function in Restricted Domain. *LNAI*, pp. 126-134.

Porzel, R., and Strube, M. (2002). Towards Context-dependent Natural Language Processing in

Computational Linguistics for the New Millennium: Divergence or Synergy. *Proceedings of the International Symposium*, pp. 21-22.

Porzel, R., Zorn, H., Loos, B., and Malaka, R. (2006). Towards a separation of pragmatic knowledge and contextual information. *ECAI-06 Workshop on Contexts and Ontologies*.

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

Rzepka, R., Ge, Y., and Araki, K. (2005). Naturalness of an Utterance Based on the Automatically Retrieved Commonsense. *Proceedings of Nineteenth International Joint Conference on Artificial Intelligence*.

Sammut, C. (2001). Managing Context in a Conversational Agent. *Electronic Transactions on Artificial Intelligence*, 5(B), pp. 189-202.

Sarma, A., and Palmer, D. (2004). Context-based Speech Recognition Error Detection and Correction. *HLT-NAACL 2004: Short Papers*, pp. 85-88.

Schumaker, R., Liu, Y., Ginsburg, M., & Chen, H. (2007) Evaluating the Efficacy of a Terrorism Question Answer System: The TARA Project. *Communications of the ACM*, 50(7), pp. 74-80.

Semeraro, G., Andersen, H. H. K., Andersen, V., Lops, P., and Abbattista, F. (2003). Evaluation and Validation of a Conversational Agent Embodied in a Bookstore. *Universal Access: Theoretical Perspectives, Practice and Experience, Lecture Notes in Computer Science*, 2615, pp. 360-371.

Serridge, B. (1997). Context-Dependent Modeling in a Segment-based Speech Recognition System. *M. Eng. thesis, MIT Department of Electrical Engineering and Computer Science*, August 1997.

Shawar, B. A., and Atwell, E. (2007). Different measurements metrics to evaluate a chatbot system. *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*.

Sherwell, B. W., Gonzalez, A. J. and Nguyen, J. (2005). Contextual implementation of human problem-solving knowledge in a real-world decision support system. *Proceedings of the Conference on Behavior Representation in Modeling and Simulation*.

Stensrud, B. S., Barrett, G. C., Trinh, V. C., and Gonzalez, A. J. (2004). Context-Based Reasoning: A Revised Specification. *FLAIRS Conference 2004*.

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

Tarau, P., and Figa, E. (2004). Knowledge-Based Conversational Agents and Virtual Storytelling. *ACM Symposium on Applied Computing*, pp. 39-44.

Towhidnejad, M. (1990). Functional Conflict Resolution in Automated Knowledge Generation. *Ph. D. Thesis, University of Central Florida*.

Turunen, M., Halkulinen, J., Smith, C., Charlton, D., Zhang, L., and Cavazza, M. (2008). Physically Embodied Conversational Agents as Health and Fitness Companions. *INTERSPEECH-2008*, pp. 2466-2469.

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997). PARADISE: a framework for evaluating spoken dialogue agents. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pp.271-280.

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

Winograd, T. (1980). What does it mean to understand language? *Cognitive Science*, 4, pp. 209-241.

Yan, P., and Zheng, F. (2004). Context Directed Speech Recognition in Dialogue Systems. *International Symposium on Tonal Aspects of Languages With Emphasis on Tone Languages*, pp. 225-228.

Young, S. (1989). The MINDS system: using context and dialog to enhance speech. *Human Language Technology Conference Workshop on Speech and Natural Language*, pp. 131-136.