MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Subset Languages for Conversing with Collaborative Interface Agents

Candace L. Sidner
Clifton Forlines

TR2002-36 August 2002

Abstract

This paper reports on experiments with subjects who must learn to use a small artificially constructed subset language (of English) to interact with a conversational spoken language system. The subjects converse with a collaborative interface agent about tasks involving TV recording and schedule navigation. The subjects perform their tasks in two conditions, one with contextually appropriate help on what to say always available on the screen and those with only a help sheet that they request to see. Our experiments indicate that users can perform their tasks in either condition, but demonstrate limits in remembering the language in a subsequent session.

Accepted for presentation at Intern'l Conf. on Spoken Language Systems, 2002

Submitted August 2002.

SUBSET LANGUAGES FOR CONVERSING WITH COLLABORATIVE INTERFACE AGENTS

Candace L. Sidner and Clifton Forlines

Mitsubishi Electric Research Laboratories
201 Broadway
Cambridge, MA 02139
{sidner, forlines}@merl.com

ABSTRACT

This paper reports on experiments with subjects who must learn to use a small artificially constructed subset language (of English) to interact with a conversational spoken language system. The subjects converse with a collaborative interface agent about tasks involving TV recording and schedule navigation. The subjects perform their tasks in two conditions, one with contextually appropriate help on what to say always available on the screen and those with only a help sheet that they request to see. Our experiments indicate that users can perform their tasks in either condition, but demonstrate limits in remembering the language in a subsequent session.

1. INTRODUCTION

Conversational speech understanding systems are now ubiquitous in the speech and language community (cf. [1], [2], [10]). We have been developing collaborative interface agents, including ones with spoken conversational capabilities ([7], [8]). In our most recent agent, we have taken the approach of designing an artificial subset language (of English) for use in spoken conversation with the agent rather than collecting a large corpus of user utterances in situ and training a language model.

Our approach is indirectly advocated by Rosenfeld and by Olsen [5], [6], who have concerned themselves with Universal Speech interfaces and speech widgets using special speech languages, respectively. While their languages appear to be artificial subset languages, no research has been done to explore the demands these languages place on human users. In particular, using an artificial subset language for speech with human users raises questions about its use, learnability and memorability by people. Furthermore, variation in the size of the grammar and lexicon, the frequency of syntactic forms from the parent natural language, and the use of synonymy in the lexicon may make a subset language harder to distinguish from its associated natural language and harder to learn and remember. In this paper we discuss an experiment in which users undertook tasks for TV and VCR use in a spoken subset of English with a collaborative interface agent. We also briefly describe the technologies that underlie our collaborative interface agent.

2. SUBSET LANGUAGES

Subset languages are languages that are subsets of a naturally occurring language that are artificially constructed for language use. They differ from naturally occurring sublanguages, which occur for specialized social and work groups, such as medicine or the military, [3] or for restricted domains such as weather reports or stock quotes [4]. Natural sublanguages may have quite large grammars and lexicons or relatively simple ones, whereas this paper focuses on artificial sublanguages that have small grammars and relatively small lexicons.

Subset languages for use in spoken conversation have the advantage that the language can be easily characterized in a grammar for a speech system and hence yield very good recognition. If subset languages are learnable by people, speech systems will have wide accessibility. The disadvantages of subset languages are that users must only use them and not utterances that are outside the subset language. Users must induce membership of the utterances in the subset. For small subset languages, the possible conversations are not as natural and free flowing as a full natural language. Subset languages also require that people remember over each use which utterances are members of the subset. In sum, users must learn to convey their intentions and constraints on intentions within the boundaries of the subset and to remember these boundaries over time.

The requirements on users who learn and use subset languages are not well studied. In a conservative approach to this problem, we designed a subset of English that contains a very common set of syntactic structures of English. The language has an English lexicon of roughly 1100 items of which 1020 are names of objects in the domain, which users do not need to remember. The grammar can be described by roughly fourteen context free rules, and is dominated by the syntax rule of an imperative verb followed by a simple noun phrase or pronoun (i.e. "that" or "me"). The language was designed to make its membership boundaries easily inferable, so that users can reliably predict the utterances that are members of the subset language.

This language serves as an utterance interface language in the MERL Entertainment Center Agent (ECA) system, in which an onscreen agent collaborates with a user to record and watch TV shows and navigate a TV schedule. Conversations such as the one in figure 1 are typical of user-agent conversations. The agent's utterances are generated as English strings in the Collagen™ [8] system and output as

synthesized speech as well as presented in text form in the agent window.

Agent says "Welcome to Diamond Talk. Try saying something shown in blue."

Navigation activities

User says "Show schedule."
Agent says "Ok, what next?"
User says "Show Friday."
Agent shows schedule for Friday 10:30 A.M.
Agent says "Ok, what next?"
User says "Show 7 P M."
Agent shows schedule for Friday 7:00 P.M.
Agent says "Ok, what next?"
User says "Show the W B."
Agent scrolls to the WB.
Agent says "Ok, what next?"
User says "??? Record a program ???"
Agent says "I didn't get that. Please repeat."

Recording a program

User says "Record a program."
Agent says "What is the program title? (exactly as on the screen)."
User says "The Invisible Man."
Agent shows episodes of The Invisible Man to be broadcast.
Agent says "Should I record The Invisible Man Friday 8:00 P.M.?"
User says "No."
Agent says "What day and time do you want to record The Invisible Man?"
User says "Friday 11 P M."
Agent scrolls to The Invisible Man Friday 11:00 P.M.
Agent adds The Invisible Man Friday 11:00 P.M. to the recording schedule.
Agent says "The Invisible Man has been added to the recording schedule Friday 11:00 P.M."
Agent says "Ok, what next?"

Playing back a program

User says "Play back a program."
Agent shows index of programs already and to be recorded.
Agent says "What is the title of the recorded program? (exactly as on the screen)."
User says "The Enemy Within."
Agent says "Should I play back The Enemy Within recorded Wednesday 7:30 P.M.?"
User says "Yes."
Agent plays The Enemy Within recorded earlier.
Agent says "Ok, what next?"

Navigation activity

User says "Show schedule."
Agent shows this week's broadcast schedule.
Agent says "Ok, what next?"
User exits.

Figure 1: A conversation with the Entertainment Center agent (ECA)

Figure 2 illustrates the user and agent as they are conversing about navigation of the TV show schedule displayed in the figure. There are 999 TV show names, and 20 TV channel names (ABC, Fox, etc) in the TV schedule. There are also windows for initial welcome, displaying the recording index, simulation of watching a TV show, and showing help text.

The ECA system is built using the Collagen™ middleware [7], [8], [9]. It took approximately four man months of effort from design of the initial subset language, agent and application to fully functional implementation. The recognition engine is IBM Embedded Via Voice used with the IBM implementation of the JSAPI. Voice synthesis is performed using the synthesizer in IBM Embedded Via Voice. All other components (speech grammars, semantic interpretation, recipe libraries, etc.) were created at MERL. There are four basic types of tasks to perform with the ECA: navigating the TV schedule (9 utterance types to do this task), recording a program (1 utterance type), playing back a prerecorded program (1 utterance type), watching a currently broadcast show (1 utterance type). Utterances for cancellation of current action, starting over and getting help are also available. The agent engaged in clarification dialogues to determine any information it needed that was not given in the user request, to disambiguate times for TV shows that occurred multiple times in the schedule, and for confirmation of actions it was to undertake. A sample clarification dialogue is part of the dialogue in Figure 1.

3. EXPERIMENT

Our study was aimed at determining: (1) whether users could complete a series tasks with the Entertainment Center agent without being tutored on how to speak to the interface or the tasks performable with the interface and underlying application; (2) whether on subsequent use, users could complete a similar series of tasks without having to relearn the subset language; and (3) whether onscreen contextual help provided significant help in speeding up their learning to perform tasks in the application. We tested 21 subjects, divided into two groups, balanced for gender, age, computer experience and general English speaking ability. All users had familiarity with computers and used them in their everyday work, some as computer professionals and some as other job professionals. We did not balance for experience with speech systems because only 2 subjects had ever used a speech system. We did not pre-test for language learnability because all the tests we could devise would pre-train our subjects for using small subset languages. Experiments were performed in an office with some variable noise conditions (e.g. air conditioning, doors slamming nearby).

Both groups performed tasks with the ECA on two encounters, spaced 24 hours apart. On the first day each subject was asked to perform a series of 12 tasks, divided into 3 sections. All subjects were shown how to use a push-to-talk key and given a written set of tasks to perform using the system. They were tested after 4 tasks to be certain they had learned the basic task of that section. Learning basic tasks indicated that the subject could not only master the task, but also had mastered how to use the subset language to accomplish the task. On the second day, each subject was given a series of 9 tasks, without testing. Tasks were randomly varied for each subject in both conditions. All subjects were timed on each task. We calculated all speech errors made by user with the agent, as well as all errors made by the recognition engine in the agent during understanding the subject's utterances. All users were asked to answer a short questionnaire and participate in a short interview following day 1 and day 2.

The two groups were distinguished by the presence or absence of context sensitive persistent help. The persistent help group (PH group) on day 1 used a version of the interface that provided a description of the relevant utterances the user could say to the agent at every point in the interaction (see Figure 2

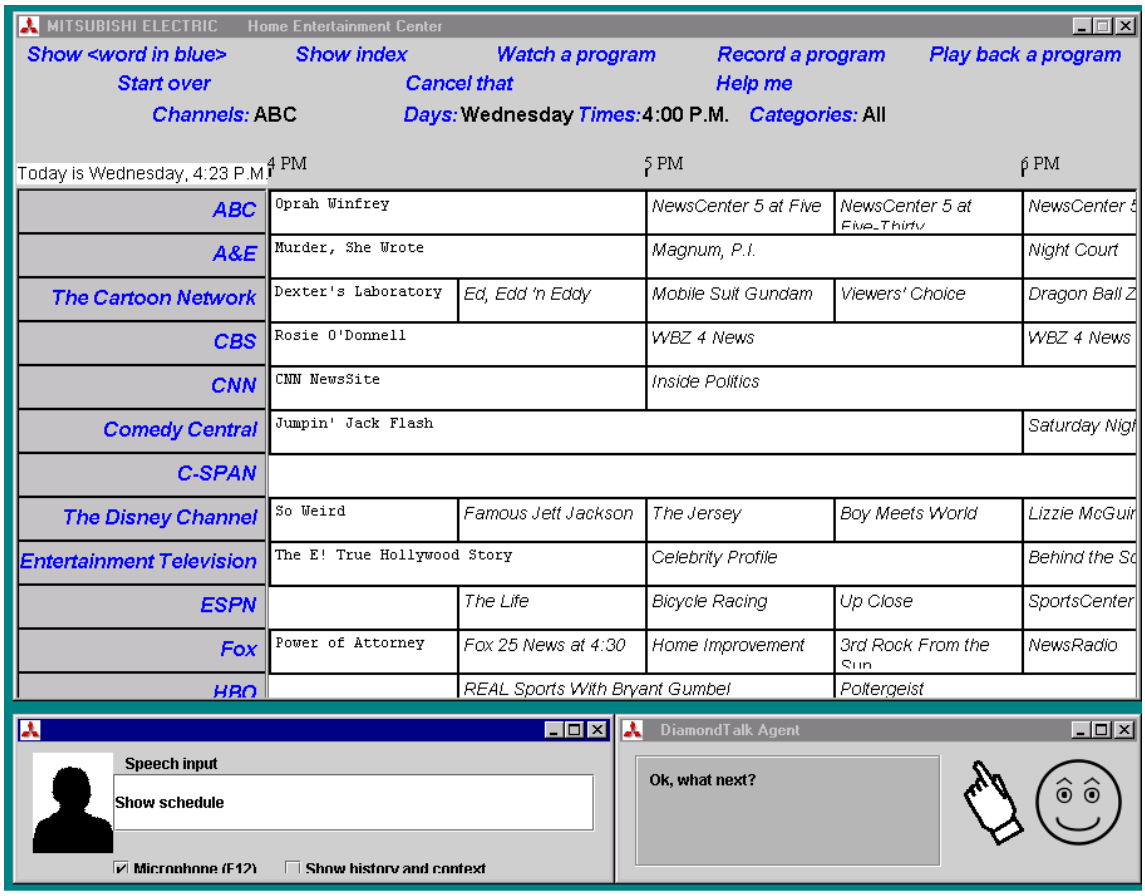

Figure 2: The Entertainment Center and Agent

for bold black words and phrases at top of the screen). The help changed with each different window that appeared in the application in performing user tasks.

The non-persistent help (NPH) group on day 1 used the interface with a blank section at the top of the screen. They were told that they could access help by saying "help me," to which a window appeared that contained descriptions of all possible utterances with a single line explanation for each utterance. On day 2, all subjects performed their tasks in the non-persistent help condition.

Among our hypotheses was a prediction that the NPH group would learn more slowly on day 1 than the persistent help group because the utterances they needed to learn in order to use the system were not as readily available. We were uncertain which group, if either, would be able to perform their tasks as quickly on day 2 because we were not sure that the subset language would be easy to remember from all of English. We thought that on day 2, the PH group would be faster than the NPH group because they had more visual experience of the subset language due to its constant onscreen presence the first day.

4. RESULTS

Both groups performed their tasks to completion, and with few exceptions, completed the tasks correctly. The mean time for the two groups to complete all tasks on both day 1 and day 2 where nearly identical (1040/1092-day 1; 389/391-day 2, mean of means). However, time turned out not to be a useful measure because variations of the task order changed the state of the application and hence the number of requests to complete a given task.

To better understand the groups' performance, we calculated the number of correct utterance requests (removing recognition errors of the speech system) on each day for each group, and number of requests for help on each day for each group. The use of a correct utterance for the given task that was appropriate for that task indicated that the subject understood and could use the relevant portion of the subset language. Requests for help (i.e. "help me") indicated that the subject had not learned or had forgotten the correct subset phrase and needed to be reminded. The PH group on day 1 was expected to use "help me" requests infrequently because of the presence of relevant utterance types on the application screen while the NPH group would use "help me" frequently because no other assistance was available.

First we present within group changes from day 1 to day 2. For the NPH group, the help ratio difference from Day 1 to Day 2 was .28 to .16 (t-test 2 tail, p=.006) while the correctness ratio difference was .60 to .75 (t-Test two tail, p=.006), a statistically significant difference. Thus this group needed less help on the second day and performed significantly

more correctly than the first day (a mean of 75% of the utterances were correctly on target for the day 2 tasks). For the PH group, the help ratio difference from Day 1 to Day 2 was .04 to .16 (t-test 2 tail, p=.011) while the correctness ratio difference was .76 to .79 (no statistically significant difference on two tail t-Test). Thus the PH group needed significantly more help on day 2 than day 1 and did not perform more correctly. Comparing across the PH and NPH groups, the NPH group performed at a difference rate across the two days of .15 while the PH group performance was -.03 (t-Test, two tail, p=.03). The NPH group compared to the PH group on help across the two days, reduced its calls for help by 0.12 while the PH group increased its calls for help by 0.12 (t-test, two tail, p=.0002). We saw no differences due to gender.

5. DISCUSSION

We offer several observations based on the above data. First, both groups initially learned the subset language well enough to perform the tasks. The PH group learned well enough to produce utterances correctly with a mean just under 80% of the time. The NPH group had less correct utterances (mean 60%), presumably because they had to learn the language from a list of utterance types that did not allow them to focus on one particular utterance type easily.

Second, in the subsequent session, the two groups perform at about the same rate. Both groups' correctness performance at day is nearly identical (.75/.79, no statistical difference), and both groups ask for help at the same rate (.16 for each group). The groups differ in the rate of change from day 1 to day 2. The NPH group improves its use of correct utterances significantly by day 2, and it uses significantly less help than it did on day 1. Third, this finding suggests that unless the groups are about to reach an asymptote in behavior, the NPH group will improve at a rate greater than the PH group. At very least, there is no advantage after the first day in onscreen help although we have weak statistical evidence from our interviews that users felt more certain about using the system with onscreen help. Persistent help may have been a good crutch to begin with, but it is a crutch that then slows its users down in future sessions.

How well have the subjects actually learned the subset language? It appears that they are a long way from perfect. Three individual subjects performed the tasks on day 2 nearly perfectly (1 perfectly in the PH group, 1 with a single help request, 1 with a single error, no help requests, both in the NPH group). This performance indicates that it is possible for subjects to correctly learn the entire language, but most of the subjects did not reach this level after 19 total tasks and 3 test tasks. Our subset language was chosen to be very minimal in terms of the size and type of grammar, the lack of synonymous terms, and the relatively small size of lexicon for words that must be actively remembered, and our tasks are few in number. Since subjects are not perfect for our subset language, we believe that users will be more challenged to learn a subset language that is larger on any of these dimensions. Additional studies with larger subset languages will be productive in determining whether our suspicions are correct.

6. CONCLUSIONS AND FUTURE WORK

Better speech and natural language understanding will ultimately eliminate the need for artificially created subset languages of the type used in the ECA. However, for current technology limitations, very simple subset languages of the type discussed here permit users to accomplish their tasks, even though they show evidence of not having learned the language completely. There appears to be no learning advantage with onscreen help, for the size of subset language and the number of tasks used in the ECA. More remains to be learned about subset languages with larger grammars, less frequently used grammatical forms, larger lexicons, and a greater number of tasks.

7. REFERENCES

1) Allen, J., G. Ferguson, A. Stent, An Architecture for More Realistic Conversational Systems, Proceedings of Intelligent User Interfaces 2001, pp. 1-8, ACM Press, New York, 2001.
2) Glass, J., E. Weinstein, SpeechBuilder: Facilitating Spoken Dialog system Development, Euro-Speech 2001, Aalborg, Denmark, vol. 2, pp 1331-1334, 2001.
3) Grishman, R. and R. Kittredge (eds.), Analyzing Language in Restricted Domain: Sublanguage Description and Processing. Erlbaum Associates, Hillsdale, N.J., 1986.
4) Kittredge, R. Variation and Homogeneity of Sublanguages, in Sublanguage: Studies of Language in Restricted Semantic Domains, R. Kittredge and J. Lehrberger (eds.), Walter de Gruyter, New York, 1982.
5) Olsen, D., Hudson, S.E., Tam, C.M., Conaty, G., Phelps, M. and Heiner, J.M. Speech Interaction with Graphical User Interfaces. Proceedings of Interact 2001, Tokyo, Japan, July 2001, IOS Press, Amsterdam, 2001.
6) Rosenfeld, R., Zhu,X. , Shriver, S., Toth, A., Lenzo, K. Black, A. Towards a Universal Speech Interface. In Proc. ICSLP 2000.
7) Rich, C., N. Lesh, C. L. Sidner . COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction, AI Magazine, Special Issue on Intelligent User Interfaces, pp. 15-25, Winter, 2001.
8) Rich, C. and Sidner, C. COLLAGEN: A Collaboration Manager for Software Interface Agents. User Modeling and User-Adapted Interaction, Vol. 8, No. 3/4, 1998, pp. 315-350.
9) Sidner, C.L., Boettner, C. and Rich, C. Building Spoken Language Collaborative Interface Agents, Lotus Technical Report TR2001-01, Lotus Development Corp. 2001.
10) Wahlster, W. Reithinger, N. Bocke, A. Smart Kom: Multi-model Communication with a Lifelike Character. Euro-Speech 2001, Aalborg, Denmark, Vol. 2. Pp. 1542-1546, 2001.