Abstract
The Web Assistant provides a conversational interface to selected sites on the World Wide Web. Users interact with the assistant by means of a flexible spoken dialogue. This paper describes the goals of the dialogue controller for the assistant and the consequent requirements for a dialogue script representation. The principle objective is to facilitate the efficient authoring of flexible conversational interactions. The assistant operates as a Web 'tour guide', and a key requirement on the dialogue controller is the ability to recognize the appropriate point at which to retake the conversational initiative from the user.

Background: Peedy the Parrot
Interaction with computers by means of a spoken dialogue is likely to become commonplace in the not too distant future. Many of the required technologies (including speech recognition, natural language processing, speech synthesis, and animation) have reached a sufficient level of maturity that the creation of an integrated conversational assistant has become feasible.

The Persona project at Microsoft Research created a limited prototype of such an assistant in 1994, called "Peedy the Parrot" (Ball et al. 1997). The Peedy prototype integrated real components (i.e. it wasn't just a mock up), but in very limited forms. The language forms that it could handle were restricted to just a few hundred variations of a dozen or so basic requests. Similarly, the task domain (selecting and playing music CD's) supported only a tiny database. These limitations, while representing significant challenges for future research, were largely a matter of scale.

A more fundamental problem with Peedy was the level of effort and expertise required to author extensions to the system. In order to add a new command to Peedy's repertoire, researchers had to:
- extend the speech recognition vocabulary and grammar to include acceptable paraphrases of the request,
- modify custom English parsing rules to ensure that those paraphrases would be recognized as variations of a single request type,
- add database queries or application commands to support the new operation,
- extend the dialogue state machine to fit the new command into a conversational sequence,
- author new 3D animations as necessary to illustrate Peedy carrying out the new request,
- write Peedy's verbal responses in the new conversational sequence,
- record any new phrases that Peedy might need to speak, along with the corresponding 'beak sync' scripts, and
- debug the entire sequence, including all interactions with existing operations.

These tasks were complex enough, and required enough specialized knowledge, that it quickly became impractical to extend the prototype to cover additional interactions. Before conversational assistants can become a practical interface technique, it must be possible for a wide variety of application designers to be able to define and extend their behaviors. Therefore, simplifying the authoring process has been a major goal of our continuing work on conversational assistants.

Web Assistant
Our current research is organized around the creation of a Web Assistant, a conversational assistant that acts as a guide to the World Wide Web. The assistant is intended to provide tours of Web sites that are relevant to the user's interests, act as a browsing companion, and handle a limited collection of navigation and search commands. The assistant's dialogue is authored in advance and downloaded to the user's machine where it controls the behavior of the character.

Our design attempts to encapsulate all of the information required by the conversational interface in a dialogue script, which is a program-like description of the possible interaction sequences that the user and assistant might traverse. This script specifies exemplars of utterances that the user might generate, as well as defining the assistant's reaction (which also depends on the current conversational state).

The input exemplars will be used by the Natural Language Processing system to generate conceptual representations of the expected inputs, which are then matched against user utterances in order to detect acceptable paraphrases. The exemplars are also used to
update the vocabulary and grammar of the speech recognizer, so that better recognition accuracy can be achieved on the most likely input utterances.

The assistant's output is presented using a general text-to-speech synthesizer and a graphical character with a limited collection of pre-authored animations. Thus the dialogue script can specify the complete conversational behavior of the assistant.

**Dialogue Management**

Within the Web Assistant scenario, there are a variety of interaction styles that the dialogue script representation should be able to support.

At one extreme, users should be able to play an almost entirely passive role, allowing the assistant to take them on a guided tour that presents information of general interest. This interaction style most resembles a broadcast medium like television, or 'Web-push' services such as PointCast.

With a small amount of interaction, users should be able to customize the guided tour to focus on sites of greater interest. The interaction might consist of answers to explicit questions from the assistant about preferences, or simple feedback about sites that are particularly interesting or that can be skipped.

At any time, users should be able to take the initiative and suspend the tour while they explore a Web site directly, perhaps using verbal commands to navigate. As they browse, the assistant may provide commentary or solicit opinions about the pages being visited. Ideally, the Web Assistant would be perceived as a pleasant companion that helps to make the browsing experience a more sociable and enjoyable one.

Finally, users may wish to make direct requests of the assistant, within the Web browsing domain. Requests for specific types of information might be anticipated in the dialogue script, enabling the assistant to respond appropriately. In other cases, the assistant could serve as an interface and advisor in the use of network services, such as search engines, to help locate the desired information.

**Dialogue Script Requirements**

The ease of dialogue authoring largely depends upon the appropriateness of the capabilities in the dialogue script representation. At least the following list of requirements need to be easily achievable:

- The assistant must be able to react promptly and appropriately to both verbal input and other events. A character that demonstrates awareness of many aspects of its environment is likely to be more believable.
- The timing of dialogue interactions is especially important. The script must be able to conveniently describe the desired interaction timing, and to react to the pacing of user responses.

- The author should be able to list a number of possible new dialogue initiatives and have the choice of directional at a particular time be controlled by a combination of constraints and chance.
- Verbal responses to the user will seem more natural if they include linguistic variation and stay consistent with the personality and mood of the assistant. It should be easy to specify both probabilistic and deterministic (based on context) control of output phrasing.
- Inappropriate repetition in the assistant's dialogue is especially disconcerting. The dialogue script must have an effective and convenient mechanism for detecting and avoiding it.
- During assistant 'monologues', users need to be able to control the pacing of the tour, ideally by means of the same non-linguistic signals that are used in human conversation.
- Awareness of recent conversational events is an important part of the capabilities of a convincing dialogue partner. The dialogue script needs to have access to an easily referenced conversational history in order to respond appropriately.
- Finally, the behavior of the assistant when it fails to respond successfully is very important. The script should be able to specify a sequence of responses to be generated when the user's input is totally unrecognizable or the assistant has run out of appropriate things to say.

**Control of Conversational Initiative**

The Web Assistant acts as a tour guide—suggesting sites that might be of interest to the user, and then supplying commentary as those sites are explored. Thus, the basic structure of a tour consists of:

- selection of a topic area,
- suggestions of sites within that topic, and
- exploration of a selected site;

the sequence is then repeated for a new site or a new topic area.

The initial assumption is that the assistant will take the initiative in guiding the interaction. This allows the user to remain passive and simply watch the tour in the style of a TV travelogue.

At any time, the user may make a request that interrupts the current tour context and shifts to a new topic. Whether there is also a shift of initiative can depend upon the specific request. For example, "Show me something funny" seems to suggest that the user wishes to remain in a primarily passive mode, but with the ability to occasionally steer the topic of the tour. In contrast, "Show me a list of humor sites" although similar in effect, suggests that the
user intends to retain control and direct the progress of the
dialogue more directly for a while.

At the conclusion of a user-directed interaction, it may
be appropriate for the assistant to retake the initiative and
suggest a new direction for the tour (or resume an
interrupted conversational context). However, the user
rarely signals the conclusion of an interaction explicitly, so
the assistant must be able to judge when it's appropriate to
suggest moving on.

In human interactions where one person is 'giving a
tour', the guide normally depends on a range of non-
linguistic signals to detect when the other's interest is
lagging and it's time to move on. Computer assistants are
likely to have to rely on only timing and verbal or mouse
input to make the same judgement. In recent experiments
where a human in another room played the role of the Web
Assistant (restricted to just verbal communication), we
found that it was often difficult for the human assistant to
detect when the user was ready to proceed to a new site.

If the user stays on one page without doing anything for
a prolonged period, the assistant can tentatively suggest a
new topic, but must be prepared for indications that the
user wishes to retain control. On the other hand, delaying
too long (i.e. being insensitive to those non-verbal signals
of impatience) can result in an especially unsatisfactory
experience, so the assistant needs to be fairly quick to
propose a new topic.

Improved techniques for managing this initiative
transition would have considerable beneficial effects on the
perceived quality of interactions with the Web Assistant.
Adjustments based on non-verbal behaviors (direct
browsing, mouse motion, image input) and a more detailed
model of the user (topic preferences, past browsing
behavior) are candidates for future investigation.

References

Persona Project at Microsoft. In Software Agents (eds.