Multilingual Conversational System Research
9807-11

Proposal for 1998-1999 Funding

Jim Glass and Stephanie Seneff

1. Introduction

Since 1989, researchers of the Spoken Language Systems (SLS) Group at the MIT Laboratory for Computer Science (LCS) have been conducting research leading to the development of conversational interfaces. A conversational interface is intended to help people access and manage information anytime and anywhere. Its realization requires the development and integration of several human language technologies, including speech recognition/synthesis and language understanding/generation [1].

All the recent conversational interfaces that were developed in the SLS Group utilize an architecture called GALAXY [2, 3] to enable the user to access information on the Internet using spoken dialogue. GALAXY differs from current speech-based interfaces in several important ways. First, it can carry on a conversation with the user in order to help him/her solve real problems.

Second, it is distributed and decentralized. GALAXY uses a client/server architecture to allow sharing of computationally expensive processes (such as large vocabulary speech recognition), as well as knowledge intensive processes. Third, it is multi-domain, intended to provide access to a wide variety of information sources and services while insulating the user from the details of database location and format. Finally, it is extensible; new knowledge domain servers can be added to the system incrementally. Over the past few years, GALAXY has been developed for several domains including travel planning, on-line shopping, and weather, making use of many on-line databases, most of them available on the Web. Users can query the system in natural English (e.g., "What flights are available from San Francisco to Osaka on United today?" "How many hotels are there in Boston?" „Will it rain tomorrow in Tokyo?‰ "Do you have any information on MIT?", etc.), and receive verbal and visual responses. The system can be accessed anywhere in the world via a Java-enabled Web Browser and a telephone. Ultimately, we envision that the interface can simply be a telephone and a cable television, thus enabling mobile and affordable information access. In fact, a telephone-only system for accessing weather information, called Jupiter, came into being in the
Spring of 1997.

Jupiter is a telephone-only conversational interface for weather information for more than 500 cities worldwide [4]. To obtain weather information, a user simply picks up the phone and conducts a verbal dialogue with the computer. The weather information is obtained from four on-line sources on the Web, and is updated several times daily. Jupiter has been available to the general public via a toll free number since June, 1997. Even with only word-of-mouth advertising, we have been able to collect nearly 50,000 sentences from more than 7,000 users. Currently, Jupiter fails to understand about one out of five within-vocabulary queries from naive users, although with correction dialogues, users usually can obtain the desired information through persistence. We have also incorporated a confidence measure, so that out-of-domain queries can be rejected.

2. Proposed Research

The objective of the proposed research is to develop the necessary human language technologies that will enable us to port our conversational interfaces to Japanese. The specific goals are to develop a version of the Jupiter system for Japanese through close collaboration with NTT researchers both in Japan and at MIT.

During the past ten years, our group has worked from time to time on conversational interfaces for Japanese [5]. However, progress has been heavily dependent on the availability of linguistic expertise. The overall goal of our proposed research is to develop Japanese conversational interfaces over the telephone for narrow domains such as access of weather information. A number of research topics that we intend to address are listed below:

Speech Recognition: How can we modify our SUMMIT recognition system, originally designed for English, to perform effectively for Japanese? In particular, what acoustic features should be used to reliably detect the accented vowels, and what sub-word units and language models would be appropriate for Japanese? Could a recognizer developed by NTT researchers be easily ported to the GALAXY architecture and the Jupiter domain?

Language Understanding: Can we adopt our English-based parsing strategy for Japanese? Should we separate the task of tokenization from the task of parsing, or should these be combined into a common framework? Should words be the basic units? Or syllables? Can we stay with a top-down parser, such as TINA, or should we consider bottom-up processing? Will the phrase structure rules of Japanese fall within the framework of our existing system?

Language Generation: What changes will we need to make to our generation system, GENESIS, to accommodate Japanese? Can we handle topicalization in the current framework?

Speech Synthesis: For English and other European languages, we have thus far relied on commercially available systems for speech synthesis. We would like to explore how a

corresponding system for Japanese can be acquired. For narrow domains such as Jupiter, we may explore the development of speech output capability based on concatenated synthesis.

Content Understanding: Another problem that we must address concerns the content, i.e., the weather information itself. We currently obtain the weather information from English web sites. In this scenario, the content must be understood and translated into Japanese. Alternatively, we could explore the feasibility of utilizing sources of information in Japan. In this scenario, the weather information must be understood in Japanese.

3. Methods and Procedures

Our approach to developing multilingual capabilities is predicated on the assumption that it is possible to extract a common, language-independent semantic representation from the input, similar to the interlingua approach to machine translation [5]. While such an approach may not be effective for unconstrained machine translation, we suspect that it is viable for conversational interfaces operating in restricted domains, since the input queries will be goal-oriented and therefore more constrained. In addition, the semantic representation may not need to capture all the nuances associated with human-human communication, since one of the participants in the conversation is a computer. Thus far, we have applied this formalism successfully across several languages and domains.

To develop a multilingual capability for our spoken language systems, we adopted the strategy of requiring that each component in the system be as language transparent as possible. The system manager, discourse component, and the database are all structured so as to be independent of the input or output language. In fact, the input and output languages are completely independent from each other so that a user could speak in one language and have the system respond in another. In addition, since contextual information is stored in a language independent form, linguistic references to objects in focus can be generated based on the output language of the current query. This means that a user can carry on a dialogue in mixed languages, with the system producing the appropriate responses to each query.

Where language-dependent information is required, we have attempted to isolate it in the form of external models, tables, or rules, for the speech recognition, language understanding, and generation components. If we are to attain a multilingual capability within a single system framework, the task of porting to a new language should involve only adapting existing tables or models, without requiring any modification of the individual components.

Another crucial aspect of multilingual system development is the collection of speech data from naïve, native speakers of Japanese. We expect to work closely with our collaborators from NTT to collect a large amount of data in Japan.

4. Budget and Requirement

This project can only go forward if we can form close collaboration with researchers at NTT. In particular, we would like to host a visiting researcher from NTT who is both a native speaker of Japanese, extremely fluent in English, and is familiar with aspects of conversational system technology. Financial support is only necessary to support such collaboration (administrative, logistic, travel, etc), and no cost for salaries per se is requested.

5. Biographical Sketches of Investigators

The proposed research will be conducted by researchers of the SLS Group at LCS. Short biographical information for the key personnel is given below. They will be assisted by other researchers and students.

James Glass received the Doctor of Philosophy degree in Electrical Engineering and Computer Science from MIT in 1988. Since then he has been at LCS where he is currently a Principal Research Scientist and Associate Head of the SLS Group. Dr. Glass has actively participated in ARPA sponsored research for over ten years in the areas of speech recognition and understanding. His interests include signal representation, pattern classification, acoustic-phonetic modelling, speech synthesis, lexical representation/access, language modelling and generation, semantic representation, and discourse and dialogue. In addition to publishing extensively in these areas, he has supervised students, and taught courses.

Stephanie Seneff received the Doctor of Philosophy degree in Electrical Engineering and Computer Science from MIT in 1985. During the 1970's, she was a member of the research staff at MIT Lincoln Laboratory, where her research encompassed a wide area of speech processing topics, including speech synthesis, voice encoding, speech recognition, and seismic signal analysis. She is currently a Principal Research Scientist in the SLS Group at LCS. Her work over the past ten years has been focused on the application of auditory modelling to computer speech recognition, and on natural language processing, including speech understanding and discourse and dialogue modelling.

6. Literature Citations

[1] Zue, V. „Conversational Interfaces: Advances and Challenges,‰ Proc. EUROSPEECH, 1997

[2] Goddeau, D., Brill, E., Glass, J., Pao, C., Phillips, M., Polifroni, J. Seneff, S. and Zue, V. „GALAXY: A Human-Language Interface to On-Line Travel Information,‰ Proc. ICSLP, 1994

[3] Lau, R., Flammia, G., Pao, C., and Zue, V. „WebGalaxy - Integrating Spoken Language and Hypertext Navigation,‰ Proc. EUROSPEECH, 1997

[4] Zue, V. , Seneff, S., Glass, J., Hetherington, L., Hurley, E., Meng, H., Pao, C., Polifroni, J., Schloming, R., and Schmid, P. „From Interface to Content: Translingual Access and Delivery of On-Line Information," ,‰ Proc. EUROSPEECH, 1997

[5] Glass, J., Flammia, G., Goodine, D. Phillips, M., Polifroni, J. Sakai, S., Seneff, S., and Zue, V. „Multilingual Spoken-language Understanding in the MIT Voyager System,‰ Speech Communication, vol. 17, 1-18, 1995