Research and Development of Multi-lingual, Multi-modal Conversational Interfaces
MIT2001-005
Jim Glass and Stephanie Seneff

MIT Principal Investigators: Drs. Jim Glass and Stephanie Seneff
NTT Principal Investigators: Drs. Kiyoaki Aikawa and Mikio Nakano

1. Background

The first phase of the NTT-MIT collaboration on research and development of multilingual conversational systems explored a language independent approach, in which a common, language-independent semantic representation is used for English and Japanese. To demonstrate the viability of this approach, we developed a Japanese version of our Jupiter weather information system, called Mokusei. The specific contributions of the project include:

1. Technology: We used the "trace" mechanism of our language understanding system to increase the efficiency in parsing Japanese sentences, which is a left recursive language. We also greatly improved our multi-lingual language generation capability by introducing a new language generation module called Genesis-II. Finally, we developed a preliminary version of a system that is capable of responding to English and Japanese appropriately, without having to perform explicit language identification.

2. System: We developed Mokusei, and delivered two complete systems to NTT in Japan. The system achieved a concept error rate of 12.6%, which is quite similar to the performance of Jupiter in its early stages of development.

3. Infrastructure: With the help of our NTT collaborators, we collected a significant amount of spontaneous Japanese data (>10,000 sentences from more than 700 calls).

4. Collaboration: Mokusei is a collaborative effort. During the period of this project, two researchers from NTT spent an extended period of time at MIT, working side by side with us. We have also written three joint papers.

While we at MIT have successfully developed several conversational systems capable of carrying on a mixed-initiative dialogue with users, building such systems requires that the system developers have intimate familiarity with the underlying human language technologies and how they interact. To overcome this limitation, we have recently started an effort in developing a utility called SpeechBuilder, which will make it easier for novice and experienced system developers to rapidly prototype new mixed-initiative conversational systems. The development of SpeechBuilder to accommodate multi-lingual and multi-modal usage constitutes a significant emphasis of this phase of our proposal research.

2. Proposed Research

Our proposed research falls into three categories: 1) infrastructures and utilities, 2) technology development, and 3) demonstration systems.

2.1 Infrastructure and Utilities

A significant portion of our effort will be devoted to the continuing design and implementation of SpeechBuilder. For example, we need to incorporate the discourse and dialogue modeling components, so that the resulting systems will be able to handle mixed initiative dialogues in a sophisticated way. We will also develop the necessary plug-and-play infrastructure, such that human language technology (HLT) components from MIT and NTT can be compared and mixed. Thus, SpeechBuilder must be able to handle HLTs developed at NTT as well as at MIT.

Two particular emphases of our effort are to extend SpeechBuilder's capabilities such that it can handle multi-lingual and multi-modal interactions. With regard to the former, we will primarily be focused on English and Japanese, with the possibility of introducing a third language, such as Mandarin Chinese or Spanish. With regard to multi-modal inputs, we will primarily be interested in the integration of speech with pointing, clicking, typing, and graphical interfaces. The objective of this part of our proposed research is to allow novice system developers to rapidly develop conversational systems simply by specifying the input sentence patterns (or actual example sentences) and providing an application backend (e.g., a database).

2.2 Technology development

SpeechBuilder will enable a greater number of system developers to build applications. The usefulness of the resulting systems, however, will ultimately be determined by the performance of the underlying HLT components. Therefore, we propose to investigate a number of new topics, including the following:

- Concatenative Speech Synthesis: With the help of our NTT collaborators, we propose to develop a Japanese counterpart to our concatenative synthesis system called Envoice. We will compare its performance to the NTT Fluet text-to-speech system.
- New Words: A conversational system will invariably encounter words that are outside of its working vocabulary. We will develop algorithms for detecting the presence of unknown words, and for gracefully incorporating them into the system.
- Methodology for Modality Integration: An effective multi-modal conversational system must be able to help the user utilize the most appropriate modality for a given situation while maintaining its linguistic competence (e.g., seamlessly transfer indirect referencing from using pronouns in speech, to using gestures to point), and to present the information appropriately (e.g., relying on visual presentation when the environment is noisy). We will investigate methodologies to properly integrate the various modalities.
- Detection and Generation of Paralinguistic Cues: Current conversational systems ignore paralinguistic cues (e.g., emotion and stress) in the speech signal, so that they are unable to detect user frustration, anger, or urgency. In addition, the responses generated by the system are always even-tempered, which may be inappropriate in certain settings. We plan to investigate acoustic cues that signify paralinguistic events, so that we may detect them in users' speech, and incorporate them into computer responses.

In addition, we also plan to collaborate with NTT researchers on benchmarking and enhancement of the NTT HLT components. An example of this effort would be the restructuring of the NTT recognizer into a finite-state transducer (FST) framework.

2.3 Demonstration Systems

There are several types of demonstration systems that we intend to develop. First, we will continue to improve Mokusei, so that the system can reach a level of performance sufficient for public usage. We will also develop conversational systems that can simultaneously handle several languages. Next, we will also develop several applications to demonstrate the effectiveness of the SpeechBuilder utility. Finally, we plan to investigate the possibility of developing bilingual systems that can help foreign language learning.

3. Collaboration with NTT

The research we propose is relevant to NTT's Humanoid Project, which pursues new HLTs to be used in real world applications. We hope that the outcome of this collaboration will contribute to the success of the NTT Humanoid Project.

We expect to collaborate with several of the working groups of the NTT Humanoid Project listed below:
WG1: Speech recognition and synthesis (headed by Shigeru Katagiri)
WG2: Dialogue processing (headed by Kiyoaki Aikawa)
WG3: Knowledge processing (headed by Shigeru Katagiri)
WG4: Communication environment understanding by visual information processing (headed by Dr. Mukawa)
WG5: System architecture (headed by Kiyoaki Aikawa, this collaboration project)

4. Period of Performance and Milestones

The project will start on July 1, 2001, and will continue for three years. While the first two years will be administered under the current NTT-MIT collaborative agreement, it is anticipated that the third year will be covered by a new, perhaps similar agreement.

4.1 First year (July 2001 - June 2002)

The first version of the multilingual SpeechBuilder toolkit will be designed and implemented which makes use of both MIT and NTT HLT components for Japanese. An application will be built using this toolkit. This application system is expected to be exhibited at the NTT CS Labs Open House to be held in June 2002. Research on corpus-based speech synthesis will begin, as will more flexible dialogue control strategies. Enhancement of some NTT-developed technologies will start (e.g., FST-based recognition), as will research on emotion recognition and language learning assistance.

4.2 Second year (July 2002 - June 2003)

The first version of the toolkit will be evaluated and reported at an international conference and through a journal paper. The second version will be designed and implemented. It will have improved understanding, generation, dialogue, and synthesis abilities and allow multi-modal input/output. By integrating this version and modules by NTT Humanoid project WG1 (speech), WG2 (dialogue), and WG4 (vision), new prototype applications will be built. This will be exhibited at the NTT CS Labs Open House in June 2003. Research on speech synthesis, dialogue control, emotion recognition and language learning will continue.

4.3 Third year (July 2003 - June 2004)

The second version of the toolkit will be evaluated and reported at an international conference and through a journal paper. The final version will be designed and implemented. Emotion recognition will be incorporated. An experimental system built using this toolkit will be exhibited in NTT CS Labs Open House and released to the press. A language learning assistance system will be built and evaluated.