Research and Development of Multi-Lingual and Multi-Modal Conversational Interfaces MIT2001-05

Progress Report: July 1, 2001—December 31, 2001

James Glass and Stephanie Seneff

Project Overview

The long-term goals of this research project are to foster collaboration between MIT and NTT speech and language researchers and to develop improved technology for natural conversational interfaces. One of the focuses of this research will be to develop methods to facilitate the development of multilingual and multi-modal conversational systems. The SpeechBuilder utility will be used as the basis for incorporating this work, which will involve the close collaboration with NTT researchers both in Japan and at MIT.

Progress Through December 2001

Over the last six months we have continued to collect data for the Mokusei weather information system, which was developed during the first phase of our NTT-MIT collaboration. We developed a bilingual capability by fusing our English and Japanese weather systems. We also began to develop a Japanese capability for our SpeechBuilder utility. The following sections describe our activities in more detail.

Mokusei Data Collection

The Mokusei weather information system is currently accessible via toll-free numbers in both Japan and the U.S. The NTT-based Mokusei provides weather information for about 150 cities in Japan, while the MIT-based system can answer about 500 cities worldwide including 40 cities in Japan. The data that we gather for Mokusei will be very valuable for creating robust telephone-based acoustic models. It will also be useful for exploring the use of language-independent acoustic models. Since the systems were made publicly available, we have collected about 2,000 user utterances from 400 calls. More than half of the utterances appear to be from naive users. Most of those data have been transcribed.

Bi-Lingual Conversational Systems

In order to expand our multi-lingual research efforts, we have explored the possibility of simultaneously performing language identification and speech recognition. This will allow users to speak to a multi-lingual system without having to explicitly pre-specify what language they wish to speak. Towards this end we have utilized the

finite state transducer (FST) recognition framework to explore the issues of combining multiple recognizers in parallel within a single search network. Using this approach, the FST search can initially consider hypotheses in multiple languages but can quickly prune away the networks of poorly scoring languages.

To evaluate our multi-lingual recognition approach, we have constructed a bilingual recognizer capable of handling weather domain queries in either English or Japanese. For English we use the recognizer from the Jupiter weather information system. For Japanese, we use the recognizer from Mokusei. In experiments, we were able to pass individual utterances through the combined bilingual recognizer in real-time, with relative word error rate increase of only 4% (from 9.3% to 9.7%). The small degradation in word error rate was due to language identification errors. The language identification error rate for this experiment was 1.25%, with a majority of the language identification errors occurring on utterances containing only one or two words.

After verifying that simultaneous language identification and recognition was both accurate and efficient, we then successfully integrated the bilingual recognizer into a complete bilingual conversational system for the weather domain. A user can speak to the system in either Japanese or English, and the system responds with an answer to the user's query in the same language that was spoken by the user.

Japanese SpeechBuilder Development

Over the last six months, we have started to augment the SpeechBuilder utility to support Japanese speech-based applications. One of the first modifications made was to enable the utility to process Unicode characters internally, and to pass them to an external interface. Currently, application developers can use Kana-kanji sequences to specify keywords and sentence patterns. Pronunciations are generated using a Japanese morphological analyzer. The speech recognizer and synthesizer are the same as those used in the Mokusei system.

The current Japanese version of SpeechBuilder has been used to create two example applications, a flight information system and telephone directory system. The SpeechBuilder developer interface is currently accessible from NTT's dialogue understanding systems group where researchers are also developing prototype Japanese applications.

Research Plan for the Next Six Months

In the coming months we plan to continue developing the Japanese SpeechBuilder capability. We plan to expand the dialogue management capability, so that more sophisticated dialogues can be configured within SpeechBuilder. By the end of the first year we would like to be able to demonstrate that a weather information system such as Mokusei could be created with the SpeechBuilder system. Finally, we plan to start developing a Japanese speech synthesis capability using the MIT corpus-based speech synthesizer. The synthesizer will initially be tailored to constrained SpeechBuilder domains.