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

Progress Report: January 1, 2002—June 30, 2002

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 June 2002

Over the last six months we have continued to develop our Japanese SpeechBuilder capability. As part of this work we have begun transferring our Japanese SpeechBuilder capability to NTT. We have also performed new work on language model learning, which will be useful for unsupervised performance improvement. Finally, we have begun to evaluate NTT dialogue control strategies in the context of the MIT Galaxy architecture. The following sections describe our activities in more detail.

SpeechBuilder Infrastructure

Over the last six months we have begun to establish a Japanese SpeechBuilder presence at NTT so that NTT researchers can create and deploy their own conversational systems. As part of this work, we set up a Galaxy distribution on a Linux PC at NTT that could be used to run SpeechBuilder applications. NTT researchers used SpeechBuilder to create several applications such as a bus timetable information system and a weather information system. It took only a couple of days to build these systems, which shows how easy the spoken dialogue system development using SpeechBuilder is. SpeechBuilder and these applications were demonstrated at the NTT Communication Science Laboratories Open House that was held on June 6 and 7. As part of the demonstration, a system that tells the time and location of the exhibitions and lectures of the open house was also demonstrated.

Language Model Learning

The creation of robust conversational interfaces is typically an iterative process whereby an initial prototype system is used to collect data. These data are typically manually transcribed and are then used to improve the acoustic and language models of the system. For systems which are deployed in a consistent acoustic environment (e.g., a telephone), it is easier to create robust domain-independent acoustic models than it is to create robust language models. Thus, we have been interested in developing unsupervised techniques for improving language models from large amounts of untranscribed data. In our initial work we have investigated a scenario where there are a small number of transcribed and a large number of untranscribed utterances available for training. This is a realistic situation that occurs soon after a prototype system has been deployed. For training, the recognition hypotheses for untranscribed utterances are classified according to their confidence

scores such that hypotheses with high confidence are used to enhance language model training. The utterances that receive low confidence can be ignored, or scheduled to be manually transcribed first to improve the language model (which can be considered as a kind of active learning). We have conducted experiments using automatic transcription of the untranscribed user utterances collected by Mokusei, a Japanese weather information system as well as Mercury, an English flight travel planning assistant system. The results have shown that the proposed methods are effective in achieving improvements in recognition accuracy while reducing the effort required from manual transcription. These methods are expected to be incorporated into future versions of SpeechBuilder.

Dialogue Modelling

As part of the NTT-MIT collaboration in the area of conversational interfaces, we have begun to explore the use of the NTT dialogue manager as part of the MIT Galaxy architecture. Over the past few months we have begun to evaluate the NTT dialogue control method called the dual-cost method using MIT human language technology. The dual-cost method enables a system to carry out an efficient dialogue according to the speech recognition accuracy and the contents of the system's database. In this preliminary work, the dual-cost method was incorporated into a spoken dialogue system generated by SpeechBuilder. The system can perform a weather information retrieval in either Japanese or English spoken dialogue.

Research Plan for the Next Six Months

In the coming months we plan to continue porting Japanese SpeechBuilder technology to NTT. As part of this work we expect to set up a mirror SpeechBuilder web site at NTT so that NTT researchers can use it without accessing the web server at MIT. We will continue to explore the NTT dialogue manager, and will examine how the performance of the dual-cost method is enhanced by estimating the speech recognition accuracy via user dialogues. Finally, we have begun research on natural sounding Japanese synthesis using the MIT corpus-based speech synthesizer.