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

Progress Report: July 1, 2002—December 31, 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 December 2002

Over the last six months we have continued to develop our Japanese SpeechBuilder capability. As part of this work we have begun to develop a more generic dialogue manager that can be more easily configured to new application domains. We have also redesigned the SpeechBuilder compiler to make the utility more portable. In addition we have begun work on natural sounding Japanese speech synthesis. The following sections describe our activities in more detail.

Dialogue Modelling

Our recent activities with SpeechBuilder have focused on making the dialogue module more easily configurable for new application domains. Dialogue management has traditionally resisted the push towards portability and rapid configurability; its role in planning and response generation had been considered too domain-dependent. In the course of building systems, however, we have noticed that basic functionalities are applicable throughout all domains. For instance, each system must gather information from a user and prompt the user for critical missing pieces, and each system must have a way of filtering responses from the database to match user-specified constraints. Furthermore, certain categories of information, such as dates and times, recur in multiple domains. Users can ask for flights on "Tuesday," or about the weather "the day after tomorrow," or for the estimated landing time of a flight scheduled for "late this afternoon."

Our approach to developing a more generic dialogue manager has been to develop a suite of self-contained dialogue flow functions that can be tailored, via an external text file, in accordance with the specifics of the application. We have also developed grammars catered to semantic concepts such as dates, times, and prices, along with a server that interprets and canonicalizes these concepts. These grammars can be used by developers in the SpeechBuilder framework to obtain precompiled meaning representations for commonly used concepts. With these semantic concepts available for any application, the work a developer must configure a conversational system can be greatly reduced. The generic dialogue manager and functions supporting common semantic concepts have been applied to several new domains, especially a hotel information domain.

SpeechBuilder Infrastructure

In order to make the SpeechBuilder utility more portable to other locations, we have been re-modularizing SpeechBuilder so that it has two distinct components: the Web-based visual development environment, and a stand-alone compiler. The front-end of the compiler converts the current XML-based domain description into an abstract syntax tree (AST). It would not be difficult to replace this front end with one that worked from a more user-friendly representation. The back-end is the most complex component. It works from the AST to generate vocabularies, grammars, catalogs, and the glue that ties them all together as a Galaxy system.

Corpus-based Speech Synthesis

Over the past six months we have started research on natural sounding Japanese speech synthesis using a corpus-based approach. Our approach has been to begin to develop a corpus-based synthesizer for our Mokusei weather information domain that focuses on naturalness within a limited domain. In addition, we are exploring longer term research issues for general synthesis. Since we have observed that perceived naturalness of Japanese synthetic speech depends strongly on the naturalness of the intonation pattern, we have begun to investigate approaches to generating an arbitrary fundamental frequency (F0) contour from a speech corpus.

We are currently exploring a method of modelling F0 by non-linear regression from the corpus, based on a statistical learning framework called Generalized Additive Models. The additive model of the F0 contour consists of intonation-phrase level F0 functions and accentual-phrase level F0 functions. Functions at both levels are jointly estimated by an iterative optimization algorithm applied to the corpus. We believe that this approach is advantageous in that it yields smooth contours and attains more robust estimation than other data-driven approaches, such as decision-tree based methods.

Research Plan for the Next Six Months

In the coming months we plan to continue development of our generic dialogue manager within the SpeechBuilder framework. We expect to complete our redesign of the SpeechBuilder compiler so that it can reside independently at NTT. We also plan to incorporate into the SpeechBuilder infrastructure additional features that researchers have found useful for their applications. These features include additional information from the recognizer such as confidence scores, and timing information for semantic concepts. We also plan to continue our development of a corpus-based Japanese speech synthesizer for the Mokusei domain, as well as more general intonation modelling for Japanese speech synthesis.