Building Japanese Conversational Systems based on the Galaxy Architecture

Mikio Nakano†, Yasuhiro Minami, James Glass, Stephanie Seneff, and Victor Zue

Abstract

This paper reviews research on multi-lingual conversational systems conducted during the five-year collaboration between NTT and MIT. It describes the development of Mokusei, a telephone-based Japanese conversational system in the weather domain, and Japanese SpeechBuilder, a toolkit that enables non-experts to build Japanese conversational systems. Both are based on the Galaxy conversational systems architecture. The experience gained by developing these systems demonstrates that the Galaxy architecture is a viable framework for multi-lingual conversational system research.

1. Introduction

Conversational systems, also called spoken dialogue systems, are computer systems that communicate with humans by oral conversation. They are expected to play a crucial role in future computer-human systems. Recent advances in speech and language processing technologies have turned this dream into reality. Most of the commercially available conversational systems such as voice-portal services adopt directed-dialogue strategies in which the system tightly controls the dialogue by asking questions that constrain the user to answer in short phrases. On the other hand, conversational systems that can understand less restricted user utterances, which could consist of dozens of words, are currently under investigation, and their performance has significantly improved over the past few years [1], [2].

One of the next challenges in conversational system research is to enable such systems to handle multiple languages. The primary objective of our project in the NTT-MIT collaboration was to find a way to build such multi-lingual conversational systems. To achieve this goal, we first ported an English-based system (Jupiter) to Japanese and developed Mokusei.

† NTT Communication Science Laboratories
Atsugi-shi, 243-0198 Japan
E-mail: nakano@atom.brl.ntt.co.jp

Our development experience showed that the Galaxy-II architecture [3], on which Jupiter is based, is a viable framework for multi-lingual conversational system research.

Although the development was successful, it required a large effort involving many spoken language processing experts. This means that the technologies developed in building Mokusei cannot be easily ported to other domains. Therefore, in the second phase of the collaborative project, we decided to build tools for facilitating multi-lingual conversational system development.

2. Mokusei: a Japanese telephone-based conversational system in the weather domain

2.1 Basic architecture

In 1999, we began development of Mokusei based on the English weather information system, Jupiter, developed at the MIT Spoken Language Systems Group. Jupiter had been available to the public via a toll-free number in the United States since May 1997. During the first four years of its deployment, over 600,000 utterances were collected from over 100,000 calls, which provide a rich corpus for training and refining system capabilities. Since Jupiter was the most mature conversational system at the beginning of our collaboration, it became the platform for our multilingual spoken language research effort. We

spent three years developing Mokusei, a conversational system that provides weather information in Japanese over the telephone [4].

Like Jupiter, Mokusei is built on the Galaxy architecture [3]. Galaxy is a client/server architecture for integrating human language technologies to create conversational systems. It enables several servers to communicate with each other through a programmable hub. These servers are specialized for individual speech and language processing tasks such as speech recognition, language understanding, and speech synthesis. Mokusei also utilizes most of the same human language technology components as Jupiter, such as the Summit speech recognizer [5], the Tina language understanding system [6], and the Genesis language generation system [7]. Figure 1 shows our basic architecture for multi-lingual conversational systems.

2.2 Japanese-specific issues

Some modifications were necessary to handle the differences between English and Japanese. For speech recognition, alternative pronunciations of lexical items were generated using a specially crafted set of phonological rules appropriate for Japanese. For language modeling, we developed a class n-gram statistical language model having a set of 56 generic word classes created manually. For language understanding, we utilized the trace mechanism of the top-down parser, Tina, to avoid inefficiencies caused by the left recursive structure. The current grammar for Mokusei has more than 900 categories and more than 2,000 vocabulary entries. For language generation, we created about 400 generation rules, along with a generation vocabulary of about 3,000 entries. To solve the problem that fluent Japanese sentences cannot be generated by simply changing the constituent order, we utilized a powerful mechanism for controlling constituent order in our generation system, Genesis-II. For synthesis, we made use of "Fluet", a software synthesizer provided by NTT Cyber Space Laboratories [8]. The database for Mokusei is nearly identical to the one used by Jupiter except that we expanded the number of Japanese cities to 144 for Mokusei. The weather information for the expanded Japanese set is obtained from both a Web site and a commercial weather information distribution service.

2.3 Data collection

To collect data about human users using Mokusei, we set up software and hardware for data collection at both MIT and NTT Atsugi R&D Center. To date, we have collected over 713 calls from naive users of the system, resulting in 10,480 utterances. This data was collected with Mokusei running at MIT. Most calls were from Japan, and about 500 of them were made

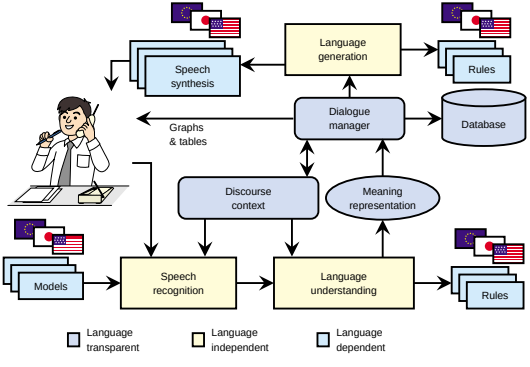

Fig. 1. Architecture of multilingual conversational interfaces.

by hired subjects who were asked to talk to Mokusei for five minutes.

When we tried to utilize the collected data for training the acoustic models, the language model for the speech recognizer, and the Tina grammar, we found that it was crucial to maintain consistency in assigning word boundaries during transcription. At first, we transcribed each user utterance as a sequence of words, each of which is written as a phoneme sequence, as in "bosutong no tengki o oshiete" ("tell me the weather in Boston"). Since there is no standard for word boundaries in Japanese, we established our own standard. As the amount of data grew, however, it became difficult to manually maintain consistent word boundaries in transcriptions. We therefore took another approach, where user utterances were transcribed phonemically, but manually segmented into *bunsetsu*. A bunsetsu is an intonational phrase consisting of a content word and zero or more function words. Although bunsetsu boundaries have not been standardized either, they can be determined far more consistently than word boundaries. We used the Tina parser to analyze morphology, segmenting each bunsetsu phoneme sequence into consistent words. In this process, variations in pronunciation, such as both 'kjaroraina' and 'karoraina' for Carolina, are reduced to a common word. The morphological analyzer correctly segments about 95% of bunsetsu in user utterances. The grammar for this morphological analyzer forms part of the grammar for sentence parsing, which guarantees consistency. We also have a program that checks consistency between the morphological analyzer and the recognition vocabulary. Using these methods, the quality of the transcription was improved, resulting in better acoustic modeling.

2.4 Evaluation

We evaluated Mokusei using a subset of the collected data. We split the naive user utterances into a training set of 542 dialogues and 8,038 utterances and a testing set of 168 dialogues and 2,442 utterances. On average there were 2.6 morae per word. The acoustic model for the speech recognizer was trained from the training set, augmented with 1,900 read speech utterances and 2,592 expert user utterances. The current recognizer has an active vocabulary of 1,151 words, with a trigram test set perplexity of 13.0. The trigram was trained only from the training set naive user utterances. On the in-vocabulary test data of 1,745 utterances containing no artifacts, the word error rate was 8.5% with a sentence error rate of 33.1% (average of 5.8 words per sentence). On the complete test set, the word and sentence error rates increased to 19.0% and 45.9%, respectively. The grammar for parsing now covers more than 75% of the naive user utterances that do not include artifacts. Overall user utterance understanding was evaluated on the key-value pair basis. For the 1,515 utterances in the test set whose transcriptions could be parsed, the concept error rate, which corresponds to the word error rate, was 12.0%. We believe that ongoing data collection using Mokusei will lead to better performance.

3. Japanese SpeechBuilder: a toolkit for building conversational systems

3.1 Motivation

Although the development and deployment of Mokusei were successful, they revealed that building a system in a new domain would take a huge amount of expert effort, which would prevent systems like Mokusei from being widely used as a form of human-computer system. If there were a tool that enabled non-expert system developers to build systems like Mokusei, we could easily apply the technologies developed through building Mokusei to other domains.

We therefore decided to build a tool set for Japanese conversational system development by integrating Japanese human language technologies in Mokusei into SpeechBuilder, a developers' toolkit that is under active development for English at the MIT Spoken Language System Group [9].

3.2 SpeechBuilder

To build a conversational system in a new domain using the Galaxy architecture, the developers need to create several kinds of knowledge sources, such as the language model for the speech recognizer and the rules for language understanding and generation. SpeechBuilder is a tool designed to make it easier to create these knowledge sources. It compiles a domain description in a format that is easy for non-experts to understand into knowledge sources that the components of the conversational systems require. The domain descriptions are in XML format, and they can be edited interactively with a Web-based system.

Two distinct methods for interfacing with SpeechBuilder are currently available. The first method begins with a list of keywords and a list of example user utterances, along with a URL address associated with a developer-defined dialogue management and response generation component. Figure 2 shows examples of keyword lists and example user utter-

keywords
city: boston, new york, chicago, ...
day: today, tomorrow, sunday, ...

example utterances
ask_weather:
tell me the weather in boston today
what is the weather like in new york
ask_rain
will it rain in chicago
tell me if it rains tomorrow

Fig. 2. Examples of keywords and sample utterances for creating domain descriptions.

ances lists, from which SpeechBuilder can create knowledge sources for speech understanding. Table 1 shows examples of the semantic representation of the user utterances that the speech understanding component outputs. Each time a user utterance is received, the user-defined dialogue management and response generation component receives its semantic representations, obtained by processing through the speech understanding module, and outputs a string that will be passed to the text-to-speech module. Typically, system developers are free to configure this dialogue management component in whatever way they like, and it is typically written in a scripting language such as Perl.

The second method for interfacing with SpeechBuilder begins with a database in the form of a table of database tuples and a set of templates for language generation, instead of the URL address. From this

Table 1. Example semantic representations derived from user utterances by the speech understanding component.

user utterance	semantic representation
tell me the weather in new york tomorrow	action=ask_weather &frame=(city=new york, day=tomorrow)
will it rain monday	action=ask_rain&frame=(day=monday)

type of description, a conversational system can be built without writing programming code, although the task domain is limited to a simple database access model. Because the Japanese version of SpeechBuilder does not yet support this second type of interface, we will not describe this aspect in detail here.

3.3 Porting SpeechBuilder to Japanese

It was necessary to make several changes to SpeechBuilder to support Japanese. First, we changed every component in SpeechBuilder to support Unicode, so that system developers could use Japanese characters in the domain descriptions. The Web-based system was also updated to support Japanese characters, as shown in Figs. 3 and 4.

The pronunciation of each word in the description was generated automatically by using the Chasen morphological analyzer [10]. We also made changes in the SpeechBuilder configuration so that it automatically selects the same speech recognizer and text-to-speech synthesizer as are used by Mokusei, whenever the developer selects Japanese as the system's language; that is, the Summit recognizer with Japanese models and Fluet are used.

Fig. 3. Web-based interface for domain descriptions in Japanese SpeechBuilder (1).

Fig. 4. Web-based interface for domain descriptions in Japanese SpeechBuilder (2).

3.4 Using Japanese SpeechBuilder in NTT Laboratories

We set up Japanese SpeechBuilder in NTT Atsugi R&D Center so that NTT researchers can use it to develop systems for human-computer dialogue collection and prototype systems for new conversational system applications. It has been used for non-experts to build systems in limited domains. In addition, it was used to build a Japanese weather information system, which was used for an experimental evaluation of NTT's dialogue control strategy [11].

3.5 Reducing the transcription costs to improve the language models

As a research topic related to SpeechBuilder, we proposed a method for reducing the effort required to transcribe user utterances to develop language models for conversational speech recognition when a small number of transcribed and a large number of untranscribed utterances are available [12]. This is a realistic situation that occurs soon after a prototype system built with SpeechBuilder has been deployed. 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 scheduled to be manually transcribed first to improve the language model. The results of experiments using automatic transcription of the untranscribed user utterances show that the proposed method is effective at improving recognition accuracy while reducing the effort required compared with manual transcription. Thus, this method will reduce the cost of creating a conversational system in a new domain. Although this method has not been incorporated into SpeechBuilder yet, we expect it will be included in a future version.

References

[1] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington, "Jupiter: A telephone-based conversational interface for weather information," IEEE Transactions on Speech and Audio Processing, 8(1):85-96, Jan. 2000.

[2] S. Seneff, "Response planning and generation in the Mercury flight reservation system," Computer Speech and Language, 16(3-4):283-312, 2002.

[3] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "GALAXY-II: A reference architecture for conversational system development," in ICSLP-98, pp. 931-934, 1998.

[4] M. Nakano, Y. Minami, S. Seneff, T. J. Hazen, D. S. Cyphers, J. Glass, J. Polifroni, and V. Zue, "Mokusei: A telephone-based Japanese conversational system in the weather domain," in Eurospeech-2001, pp. 1331-1334, 2001.

[5] J. R. Glass, "A probabilistic framework for segment-based speech recognition," Computer Speech and Language, 17:137-152, 2003.

[6] S. Seneff, "TINA: A natural language system for spoken language applications," Computational Linguistics, 18(1):61-86, 1992.

[7] L. Baptist and S. Seneff, "Genesis-II: A versatile system for language generation in conversational system applications," in ICSLP-00, pp. III:271-274, 2000.

[8] K. Hakoda, T. Hirokawa, H. Tsukada, Y. Yoshida, and H. Mizuno, "Japanese text-to-speech software based on waveform concatenation method," in AVIOS '95, pp. 65-72, 1995.

[9] J. Glass and E. Weinstein, "SpeechBuilder: Facilitating spoken dialogue system development," in Eurospeech-2001, pp. 1335-1338, 2001.

[10] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. "Morphological Analysis System ChaSen version 2.2.1 Manual," Nara Institute of Science and Technology, 2000.

[11] K. Dohsaka, N. Yasuda, and K. Aikawa, "Efficient spoken dialogue control depending on the speech recognition rate and system's database," in Eurospeech-2003, (to appear).

[12] M. Nakano and T. J. Hazen, "Using untranscribed user utterances for improving language models based on confidence scoring," in Eurospeech-2003, (to appear).

Mikio Nakano
Senior Research Scientist, Media Information Laboratory, NTT Communication Science Laboratories.
He received the M.S. degree in coordinated sciences and Sc.D. degree in information science from the University of Tokyo, Tokyo in 1990 and 1998, respectively. In 1990 he joined NTT, where he has been working on spoken dialogue systems and spoken language understanding. From 2000 to 2002, he was a visiting scientist at the Spoken Language Systems Group, MIT Laboratory for Computer Science, where he engaged in research and development of multilingual spoken dialogue systems.

Yasuhiro Minami
Senior Research Scientist, Media Information Laboratory, NTT Communication Science Laboratories.
He received the M. E. degree in electrical engineering and the Ph.D. in electrical engineering from Keio University, Kanagawa in1988 and 1991, respectively. He joined NTT in 1991. He had worked in robust speech recognition. He was a visiting researcher at MIT from 1999 to 2000. He is interested in modeling for speech recognition.

James Glass
He is a Principal Research Scientist, Head of the Spoken Language Systems Group in the MIT Computer Science and Artificial Intelligence Laboratory, and a member of the Speech and Hearing Bioscience and Technology Program of the Harvard-MIT Division of Health, Sciences, and Technology.
He received the Ph.D. degree in electrical engineering and computer science from MIT in 1988. After starting in the Speech Communication group at the MIT Research Laboratory of Electronics, he has worked since 1989 at the Laboratory for Computer Science. His primary research interests are in the area of speech communication and human-computer interaction, centered on automatic speech recognition and spoken language understanding. He has lectured, taught courses, and has many publications in these areas. He has been a member of the IEEE Acoustics, Speech, and Signal Processing, Speech Technical Committee and an associate editor for the IEEE Transactions on Speech and Audio Processing.

Stephanie Seneff
She is a Principal Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory.
She received the B.S. degree in biophysics and the M.S., E.E., and Ph.D. degrees in electrical engineering and computer science from MIT. Her research interests span a wide spectrum of topics related to conversational systems, including phonological modeling, auditory modeling, computer speech recognition, statistical language modeling, natural language understanding and generation, discourse and dialogue modeling, and prosodic analysis. She has published numerous papers in these areas, and she is currently supervising several students at both master's and doctoral levels.

Victor Zue
Professor of electrical engineering and computer science at MIT and Co-Director of the Institute's Computer Science and Artificial Intelligence Laboratory (CSAIL).
He is also the first holder of the Delta Electronics Chair endowed for senior researchers. His main research interest is in the development of spoken language interfaces to make human/computer interactions easier and more natural, and he has taught many courses, written many articles, and lectured extensively on this subject. Over the years, he and his colleagues at the Spoken Language Systems Group have pioneered the development of many systems that enable a user to interact with computers using multiple spoken languages (English, Japanese, Mandarin, and Spanish). He is a Fellow of the Acoustical Society of America.