Mokusei: A Telephone-based Japanese Conversational System in the Weather Domain

Mikio Nakano, Yasuhiro Minami, Stephanie Seneff, Timothy J. Hazen, D. Scott Cyphers, James Glass, Joseph Polifroni, and Victor Zue

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139, USA

NTT Corporation
Atsugi 243-0198, Japan
and Kyoto 619-0237, Japan

Abstract

This paper describes MOKUSEI, an end-to-end Japanese version of our JUPITER weather information system. MOKUSEI delivers weather information over the phone through natural conversation with the user. For the most part, MOKUSEI uses the same components for recognition, understanding, and generation that JUPITER uses, and the database and the semantic frames for the weather information content are also shared. However, MOKUSEI motivated us to redesign our GENESIS generation system, in order to improve the quality of translations of weather reports into Japanese. We also had to develop new ways to transcribe user utterances through morphological analysis. MOKUSEI is fully functional and has already been used for data collection with about 700 naive users. These data have been used for improvement and evaluation of MOKUSEI. This paper also presents the result of evaluating the current version of MOKUSEI.

1. Introduction

For more than a decade, the Spoken Language Systems Group has been conducting research leading to the development of conversational systems that enable users to access and manage information using spoken dialogue. While most of our systems have been developed for English, multilinguality has always been an important topic on our research agenda. Our approach to developing multilingual conversational systems is predicated on the assumption that it is possible to extract a *common* language-independent semantic representation from the input, similar to the *interlingua* approach to machine translation [1]. Whether such an approach can be effective for unconstrained machine translation remains to be seen. However, we suspect that the probability of success is high for spoken language systems operating in restricted domains, since the input queries will be goal-oriented and therefore more constrained. Thus far, we have applied this formalism successfully across several languages and domains [2, 3].

In 1997, we introduced the JUPITER weather information system in English [4]. JUPITER has been available to the public via a toll-free number in the United States since May 1997. Over the three year period since its introduction, we have collected over 400,000 utterances from over 58,000 calls, which provide a rich corpus for training and refinement of system capabilities. Since JUPITER is our most mature conversational system to date, it has become the platform for our multilingual spoken language research effort. This paper describes MOKUSEI[1], a conversational system that provides weather information in Japanese over the telephone. MOKUSEI employs the same Galaxy Communicator architecture [5] as its English predecessor. It also utilizes many of the same human language technology (HLT) components, although some modifications were necessary to account for differences between English and Japanese. This paper describes our system development effort and the performance evaluation results. Due to space limitations, readers are referred to our other publications for a background description of JUPITER. A detailed explanation of the initial development effort can be found in our earlier report [6].

2. System Architecture

The overall system consists of a number of specialized servers that communicate with one another via a central programmable hub, using the Galaxy Communicator architecture [5]. In a telephone-based configuration, an audio server captures the user's speech and transmits the waveform to the speech recognizer. The language understanding component parses the recognizer's word graph and delivers a semantic frame, encoding the meaning of the utterance, to the discourse resolution component. The resulting "frame-in-context" is transformed into a flattened E-form (electronic form) by the generation server. This E-form is delivered to the dialogue manager, and provides the settings of the dialogue state.

The dialogue manager consults a dialogue control table to decide which operations to perform, and typically engages in a module-to-module sub-dialogue to retrieve tables from the database. It prepares a response frame, containing weather reports represented as semantic frames, which is sent to the generation component for translation into the target language[2]. The speech synthesizer then translates the response text into a speech waveform, which it sends to the audio server. Finally, the audio server relays the spoken response to the user over the telephone. A detailed record of the entire dialogue, including state information, is logged along with user utterances for later examination and reprocessing.

To develop a multilingual capability for our spoken language systems, we have adopted the strategy of requiring that

This work was supported by NTT, and by DARPA under contract N66001-99-1-8904 monitored through Naval Command, Control and Ocean Surveillance Center. Drs. Nakano and Minami from NTT participated in this research at MIT as Visiting Researchers.

[1] MOKUSEI is the Japanese name for the planet Jupiter.

[2] Our English JUPITER system translates weather reports from English back into English.

to be presented at the 7th European Conference on Speech Communication and Technology (EUROSPEECH), Aarlborg, Denmark, September, 2001

each component in the system be as language neutral as possible. The dialogue management, discourse resolution, and the application back-end are all structured so as to be independent of the input or output language. In fact, the input and output languages are completely independent of each other so that a user could speak in one language and have the system respond in another. In addition, since contextual information is stored in a language independent form, linguistic references to objects in focus can be generated based on the output language of the current query. This means that a user can carry on a dialogue in mixed languages, with the system producing the appropriate responses to each query.

2.1. Speech Recognition

Speech recognition for the MOKUSEI system is performed using the SUMMIT speech recognition system [7]. Currently the recognizer uses a vocabulary of 1,151 *words* relevant to the weather domain. A majority of these words are names of geographic locations and words describing various weather conditions. A phonetic pronunciation for each word has been created using a standard set of Japanese phonetic units. For language modeling, we developed a class n-gram having a set of 56 generic word classes created by hand.

To account for phonological variations, a set of phonological rules is applied to the basic pronunciations of each word. The output is a graph of possible alternate pronunciations. For example, one set of phonological rules accounts for the deletion of /i/ and /ɯ/, which is common in Japanese. An example of this is the word sequence *desu ka* being pronounced as /d e s k ɑ/.

For the initial version of the MOKUSEI recognizer, the acoustic models were trained entirely from English utterances. These models were used to create forced transcriptions of the early sets of Japanese data that were collected. As Japanese data became available, the acoustic models were retrained using a combination of English and Japanese utterances. As more Japanese data were collected, the dependence on English data for training was eventually eliminated.

The MOKUSEI recognizer was created in a straightforward manner using the standard tools of the SUMMIT recognizer. SUMMIT did not require any design adjustments to handle Japanese. Significant effort was required only to create the vocabulary list, pronunciations, and phonological rules, as these elements could not be bootstrapped from their equivalent counterparts in the English system.

2.2. Natural Language Understanding

Once the recognizer has proposed a word graph of promising candidates, the graph is parsed by the natural language understanding (NLU) component to produce a semantic frame. For all of our research in NLU, we have made use of the TINA system [8], first developed to accommodate database query domains in English. TINA is a top-down parser which supports an automatically trainable probability model and a "trace" mechanism to handle movement, which is prevalent in English questions (compare "Is this restaurant on *Main Street*?" with "*What street* is this restaurant on <trace>?").

We were at first concerned that a top-down parser might not be an appropriate choice for Japanese, which is a left-recursive language. The problem is that the system must propose the entire parse column above the first word before it has seen the rest of the sentence. For example, the three sentences in Fig. 1 all begin with the same word, but have significantly different parse

(Nihon wa) doo desu ka?
((Nihon no) tenki wa)) doo desu ka?
(((Nihon no) Tokyo no) tenki wa) doo desu ka?

Figure 1: *Three sentences beginning with the word "Nihon," with differing roles and parse depths for this word.*

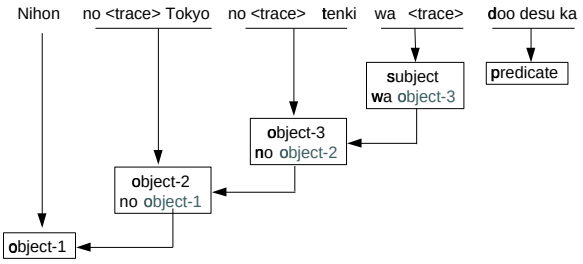

Figure 2: *Illustration of use of "trace" mechanism to efficiently parse left-recursive structures in Japanese.*

structures, as indicated by the bracketing. A computationally expensive solution is to propose all possible parse structures and let the later evidence eliminate inappropriate ones.

We found that a much better solution to this problem was available through TINA's trace mechanism. Using this approach each new content word is parsed first in a shallow parse tree, and then later moved to a position just after the subsequent particle that defines its role. The upper columns of the parse tree are not constructed until after the appropriate role has been identified. This has the intended effect of reordering the constructs to appear right-recursive. This process is illustrated in Fig. 2. The change reduced computational requirements by two orders of magnitude.

2.3. Natural Language Generation

To translate weather reports into Japanese, we made use of our GENESIS-II generation system [9]. GENESIS-II generates text in the target language from a semantic frame, obtained by parsing the weather forecast using TINA. Generation is controlled by a message file that recursively describes the ordering of nested constituents in the semantic frame. A vocabulary file provides the mappings from the semantic tags to the appropriate target language surface realization.

Often, a literal translation, even properly reordered, is inappropriate, as illustrated in Fig. 3, with the semantic frame for the original sentence shown in Fig. 4. Notice that fluent translation requires a major reorganization of the sentence structure. In particular, the word "rain" ("ame") needs to appear *between* the IN_TIME predicate ("gozenchuu") and the word "possibility" ("kanoosee"), which is impossible if each frame is required to generate its entire string as a single unit.

In part because original GENESIS system was unable to handle this kind of the situation, we have been motivated to redesign our generation system. The resulting GENESIS-II is significantly more flexible and intuitive. It provides a powerful "pull" mechanism that overcomes the above problem. Any constituent can pregenerate by name any other constituent among its grandchildren (<--), or even among its however deeply nested descendants (<==). In our example, the weaker pull (<--) is sufficient, and is accomplished through the following rule: precip_act <-- in_time ... :name ... probability

rain (possible (in the morning))	[English original]
((morning in) possible) rain	[Japanese ordering]
((gozenchuu ni) kanoo na) ame	[*nonsense*]
gozenchuu (((ame no) kanoosee ga) arimasu)	[*fluent*]
morning (((rain of) possibility) SBJ exist)	[English equivalent]

Figure 3: *An example weather phrase whose literal translation is nonsensical, along with the appropriate translation that is produced by our system.*

```
{c weather_event
   topic:
    {q precip_act name: "rain"
       pred:
        {p possibility
           qualifier: "possible"
           pred:
            {p in_time
               topic:
                {q time_of_day
                   name: "morning"
                   quantifier: "def"} } } } }
```

Figure 4: *Semantic frame for the sentence, "rain possible in the morning," showing the embedded* IN_TIME *predicate.*

resulting in the fluent translation shown in Fig. 3. GENESIS-II has another new *selector* feature, described elsewhere [6], which allows context-dependent word selection. This is essential for generating natural Japanese sentences.

2.4. Dialogue Management and Synthesis

The dialogue manager of MOKUSEI is identical to that of JUPITER except that we had it default to the Celsius scale for temperatures when the language is Japanese. This requires a mechanism to interpret frequent references to ranges of temperature, such as "high upper 60s". Aside from this one feature, the dialogue manager does not need to know either the input or output languages. It prepares a response as a semantic frame, which later gets translated into text in the target language by the generation system.

For Japanese synthesis, we made use of FLUET, a software synthesizer provided by NTT Cyber Space Labs. [10]. We compiled it with a wrapper to allow it to communicate with the other servers via the standardized protocol.

2.5. Content Processing

JUPITER updates its weather database three times a day from four Web sites, and updates information about current temperature, humidity, and pressure continuously from its satellite feed. The update involves several steps. After the new weather reports are retrieved from the Web sites, they are all parsed into semantic frames using our TINA NLU system. The frames are then examined automatically for semantic content, and indexed under all weather categories that are applicable. Finally, the relational database tables are updated.

The database for MOKUSEI is nearly identical to the one used by JUPITER. However, since JUPITER knew of only six major cities in Japan, we expanded the number of Japanese cities to 144 for MOKUSEI. We also provided information about the prefecture and region for each city in a geography table. The weather information for the expanded Japanese set is obtained from both a Web site and a commercial weather information distribution service.

3. Data Collection and Evaluation

3.1. Data Collection

Data collection is a vital part of conversational system development. Once a preliminary version of every component is available, users can talk to the system to obtain information. Later perusal of the resulting log file reveals problems that can then be repaired by system developers. The users' queries are transcribed, and the text is used to guide expansion of the grammar rules for the NLU component, as well as the language model for recognition. The acoustic models can be retrained on the collected waveforms. The software and hardware for data collection reside at both MIT and NTT's Atsugi R&D Center.

To date, we have collected over 713 calls from naive users of the system resulting in 10,480 utterances. These data were collected with MOKUSEI running at MIT. Most calls were from Japan, and about 500 of the calls were made by hired subjects who were asked to talk to MOKUSEI for five minutes. When these data were collected, MOKUSEI's performance was somewhat limited because of the lack of training data.

3.2. Transcription

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 of word boundaries for transcriptions. At first, we transcribed each user utterance as a sequence of words, each of which is written as a phoneme sequence, as in "rosangzhjerusu no tengki o oshiete (tell me the weather in Los Angeles)." Since there is no standard for word boundaries (i.e., tokenization) in Japanese, we established our own standard.

As the amount of data grew, it became difficult to manually maintain consistent word boundaries in transcriptions. We considered two approaches to this problem. One was to use Japanese morphological analyzers which take Japanese text written in kanji-kana sequences[3] as input and output morphemes. This method is often used to build n-gram language models for Japanese dictation systems from text corpora such as newspaper articles [11]. To utilize this method, utterances need to be transcribed in kanji-kana sequences. The problem with this approach is that, because there is no standard for Japanese orthography, there are many possible kanji-kana sequences for the same utterance that are semantically equivalent. This causes problems when we try to use these transcriptions to develop the recognition vocabulary and grammar for language understanding.

We therefore took another approach, where user utterances were transcribed phonemically, but automatically segmented into *bunsetsu*. A bunsetsu is an intonational phrase consisting of a content word and a number (zero or more) of function words. Although bunsetsu boundaries have not been standardized, they can be determined far more consistently than word boundaries. We use the TINA parser to analyze morphology, segmenting each bunsetsu phoneme sequence into consistent words, as illustrated in Fig. 5. In this process, variations in pronunciation, such as both 'rosangzhjerusu' and 'rosangzerusu' for Los An-

[3] A kanji is a Chinese character and a kana is a phonogram.

transcription:	getsujoobino tookjoono tengkio shiritaingdesukedo
segmented sentence:	getsujoo bi no tookjoo no tengki o shiri tai ng desu kedo (I want to know the weather in Tokyo for Monday)

Figure 5: *An example of a transcription and its morphological analysis.*

geles, are reduced to a common word. Formerly this was done manually, sometimes leading to inconsistencies. The morphological analyzer correctly segments about 95% of bunsetsu in naive and expert user utterances.

The grammar for this morphological analyzer forms a 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 improvement in the acoustic model.

3.3. Performance Evaluation

Using the collected data, we evaluated MOKUSEI. We split the naive user utterances into a training set, which comprises 542 dialogues and 8,038 utterances, and a test set of 168 dialogues and 2,442 utterances. On average there are 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 is 8.5% with a sentence error rate of 33.1% (average of 5.8 words/sentences). On the complete test set the word and sentence error rates increased to 19.0% and 45.9%, respectively. These results are similar to the performance we obtained for our English weather system [4].

The grammar for sentence parsing currently has more than 500 categories and nearly 2,000 vocabulary entries. It has been refined so that it covers more of the collected naive user utterances while avoiding a decrease in parsing speed. It now covers more than 75% of the naive user utterances that do not include artifacts.

Overall user utterance understanding was evaluated using our evaluation framework [12], which evaluates understanding by comparing key-value pairs obtained both from recognition results and transcriptions. For the 1,515 utterances in the test set whose transcriptions are parsable, concept error rate, which corresponds to word error rate, is 12.0%.

For generation, we have created about 400 generation rules, along with a generation vocabulary of about 3,000 entries. The spoken responses for the manually transcribed test set utterances were manually checked, and more than 90% of them can be considered fluent.

4. Summary

This paper describes MOKUSEI, a weather information system that enables Japanese speakers to obtain real weather information for cities worldwide by conversing with the system. We have expanded the capabilities of our GENESIS generation system such that it is now possible to generate high quality translations of the English weather reports. We have also incorporated morphological analysis of user utterance transcriptions into bunsetsu sequences to maintain word boundary consistencies. The performance of the current version is good enough to allow naive users to use it, although we believe that ongoing data collection using MOKUSEI both at MIT and in Japan will lead to better performance. Together with MUXING, the Mandarin Chinese version of JUPITER [3], MOKUSEI shows that JUPITER and its underlying GALAXY architecture are viable frameworks for multilingual conversational system research.

5. Acknowledgements

Lauren Baptist designed and implemented GENESIS-II. I. Lee Hetherington and Norihito Yasuda's help was invaluable. We would also like to acknowledge the many colleagues from NTT who have contributed to our data collection effort.

6. References

[1] W. Hutchins and H. Somers, *An Introduction to Machine Translation,* Academic Press, 1992.

[2] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual spoken-language understanding in the MIT Voyager system," *Speech Comm.*, 17:1–18, 1995.

[3] C. Wang, S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. Zue, "Muxing: A telephone-access mandarin conversational system," in *Proc. ICSLP*, 2000, pp. II:715–718.

[4] 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 Trans. Speech and Audio Proc.*, 8(1):85–96, 2000.

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

[6] V. Zue, S. Seneff, J. Polifroni, M. Nakano, Y. Minami, T. Hazaen, and J. Glass, "From JUPITER to MOKUSEI: Multilingual conversational systems in the weather domain," in *Proc. MSC*, 2000, pp. 1–6.

[7] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP*, 1996, pp. 2277–2280.

[8] S. Seneff, "TINA: A natural language system for spoken language applications," *Comp. Ling.*, 18(1):61–86, 1992.

[9] L. Baptist and S. Seneff, "GENESIS-II: A versatile system for language generation in conversational system applications," in *Proc. ICSLP*, 2000, pp. III:271–274.

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

[11] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Free software toolkit for Japanese large vocabulary continuous speech recognition," in *Proc. ICSLP*, 2000, pp. IV:476–479.

[12] J. Polifroni and S. Seneff, "Galaxy-II as an architecture for spoken dialogue evaluation," in *Proc. LREC*, 2000, pp. 725–730.