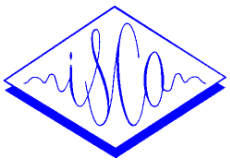

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A PORTABLE DEVELOPMENT TOOL FOR SPOKEN DIALOGUE SYSTEMS

Satoru KOGURE and Seiichi NAKAGAWA
{*kogure, nakagawa*}*@slp.tutics.tut.ac.jp*

Dept. of Information and Computer Sciences, Toyohashi Univ. of Technology
1-1 Hibarigaoka, Tempaku, Toyohashi, Aichi, 441-8580 Japan

ABSTRACT

Speech recognition and language processing technologies have recently been improved, and speech recognition systems and dialogue systems are now in practical use. Nevertheless, not only the fundamental techniques but also the techniques for improving portability and expansibility of the systems need to be developed further. Whereas we need many user utterances to construct N-gram based language models for speech recognition. It is difficult to construct the language model for a new task/domain because many utterances cannot be collected easily. So, we propose an efficient method that is able to construct a language model, preparing at least only several hundreds utterances. We designed a domain-independent platform for developing any spoken dialogue systems for retrieving information from a database, and used the platform to build a literature retrieval system. When a system is implemented, the various items in the dictionary for understanding can be semi-automatically created by preparing the databases, dialogue samples and so forth.

1. INTRODUCTION

Recently, much research has been done on the robustness and reliability of spoken dialogue systems. We developed a "Mt. Fuji sightseeing guidance" system which used touch screen input, speech input/output and graphical output, and have improved the sub-modules of a speech recognizer, natural language interpreter, response generator and multi-modal interface [1, 2, 3, 4, 5]. However, all of these modules except for the speech recognition module depended on a given task or domain.

As speech recognition systems are increasing by being used in practical applications, spoken dialogue systems will also become more widespread. However, the cost of developing a new spoken dialogue system is enormous. The systems that have been developed so far can not be transferred to other domains easily, and yet a highly-portable system that can be easily adapted to another domain or task urgently needs to be developed. There are several examples of researches that focused on high portability and expansibility [6, 7]. In [6], a prototype could be simply constructed even in a complicated speech dialogue system using the PIA system, which was implemented using Visual Basic. This system placed priority on achieving high robustness of speech recognition and high naturalness of generated dialogue. However, the system limited the task to the domain of knowledge search. The system produced experimentally by the project of REWARD (Real World Applications of Robust Dialogue) [7] allowed the development and debug going of the system to be controlled by the developer. This system attempted to implement a spoken dialogue system through a telephone line. In the CSLU Toolkit system [8] which was developed by CSLU (Center for Spoken Language Understanding) of OGI, the system can construct the dialogue's application using the speech, even if the developers do not have any knowledge of the speech dialogue system at all. Furthermore, modules such as natural language understanding, speech synthesis and animation of face images as well as speech recognition could be easily constructed. However, the parsing capability has been limited in that each component can not understand complicated grammar, for example. This CSLU system was developed to provide classroom training on speech processing. M. Sasajima also proposed a new framework for developing spoken dialogue system [9]. In this framework, dialogue control is described by a unification-based script handling instruction set language, which is similar to PASCAL. In this system, a constraint is that the speech recognition module uses keyword-spotting to understand spontaneous speech.

During system development, we took into consideration the portability of the spoken dialogue system. We developed a prototype system that retrieves information from a given database using speech input. If a database and dialogue samples of retrieving information from the database are given, a spoken dialogue system for supporting the retrieval can be easily developed. We applied the system to the task of document retrieval.

2. Highly-portable System for Information Retrieval from Database

2.1. Required data

We consider what kind of data may be prepared as task-dependent knowledge when the task is applied. In the proposed framework, the system developer prepares the following:

- A generally usable database
- The format information of each field of the database
- A corpus of user utterances (dialogue examples)

The database information retrieval system first requires a database (generally, one that is open to the public) as a retrieval object. The format information of each field in the database should be defined in order to access the database. In addition, since the dictionary and language model are adapted to the database information retrieval system in the task, a corpus of the user utterances is required.

2.2. Dictionary and knowledge generated by the framework

Using the above data set, the framework generates the following:

- Morphological dictionary
- Semantics dictionary
- Retrieval slot
- Database for PostgreSQL

In the proposed system, a morphological analysis dictionary and semantics dictionary are required, which are used when a user utterance is analyzed. We used JUMAN [10] as a morphological analysis module. A noun dictionary and a proper noun dictionary are generated using all the items of the database and user utterance corpus within the morphological analysis dictionary (default dictionary supplied with JUMAN). The format information of the database indicates to which dictionary each field of the database should be registered, and this format information is used to generate the above dictionaries. The semantics dictionary used in the semantic analysis contains case frame information for verbs and concept information for nouns. The case frame information is created from the user utterance corpus, and the concept information is created from the database and the user utterance corpus. The relational conditions between case and nouns, and the concept information of nouns are extracted from the EDR dictionary (JEDRI Ltd.), which has 250,000 Japanese words and 200,000 concepts.

2.3. Speech recognition

SPOJUS-Y [1] based on context-free grammar and SPOJUS-Z [11] based on N-gram, which we constructed, are used in the speech recognition module.

The acoustic processor samples the input speech at the rate of 12 kHz and the succeeding LPC analysis stage generates the 14 LPC cepstral coefficients which are then used to calculate the feature parameters of 10 LPC mel-cepstral coefficients and the additional dynamic and segmental features. All the experiments shown in the following sections are based on the features of 10 LPC mel-cepstral coefficients and their derivatives. The acoustic model consists of 113 Japanese syllable-based HMMs, which have 5 states, 4 Gaussian densities (full covariance matrices) and 4 discrete duration distributions.

The search algorithm employed in the system is based on the Viterbi-based one-pass decoding algorithm using the HMM-based sub word acoustic models and it is integrated with the context-free grammar parser or N-gram constraints in a frame-synchronous manner. In the former, the Earley-like parsing algorithm employed in our system was extended to generate a dynamic representation of the linguistic constraint as a finite state network. Since the number of grammatical states becomes large, we used the beam search method in the process of prediction and pruning of unreliable candidate branches from the search space.

In general, modeling and processing of phenomena such as filled pauses and restarts for a language processing system are difficult and thus such phenomena should be detected at the decoding (transcribing) level and excluded from further consideration. We proposed an unknown-word processing method for detecting out-of-vocabulary words and filled pauses in the decoding process [1, 11].

2.4. Language model

It has been proved that the recognition rate by a bigram language model made automatically from the corpus is higher than that by context-free grammar made manually, if we can prepare thousands of user utterances [12]. However, it is difficult to collect thousands of sentences for the new task. Therefore, a method for adapting the task/domain independent initial bigram using a small amount of the user utterance corpus should be considered.

When word bigram $P'(w_2|w_1)$ and class bigram $P'(c_2|c_1)$ were calculated from same training corpus where vocabulary sets are V, we must consider four pairs of bigram as follows: (1) w_1 and w_2 are known words, (2) w_1 is a known word and w_2 is an unknown word, (3) w_1 is an unknown word and w_2 is a known word, and (4) w_1 and w_2 are unknown words. The bigram probability of test corpus which consists of unknown words is calculated as the following equation.

$$
\begin{aligned}
(1)\quad P(w_2|w_1) &= P'(w_2|w_1) \\
(2)\quad P(w_2|w_1) &= P(c_2|w_1) \times P(w_2|c_2) \\
&= \sum_{w_i \in c_2} P'(w_i|w_1) \times P(w_2|c_2) \\
(3)\quad P(w_2|w_1) &= \frac{C(w_2, c_1)}{C(c_1)} \\
(4)\quad P(w_2|w_1) &= P'(c_2|c_1) \times P(w_2|c_2)
\end{aligned}
$$

3. Application to Database Document Retrieval System

The flow of the task adaptation is shown in Figure 1. Since the application developer prepares the database and user utterance corpus, the task adaptation of the system can be carried out by generating the required dictionaries automatically. In the following explanation, the document retrieval system is adopted as an example of the task for the database information retrieval systems.

3.1. Database and definition

A great number of articles were extracted from the Institute of Electronics, Information and Communication Engineers Journal of Japan, the Information Processing Society Journal of Japan, The Acoustical Society Journal of Japan and the Japanese Society for Artificial Intelligence Journal, and the resulting paper information (3092 cases) in Japanese was used as the database. The field information of the database was defined as the data type of each item of the database and the method processing the data. *Transfer format information* describes how each item of paper information is processed. And in order to register table information in the database for PostgreSQL, the ta-

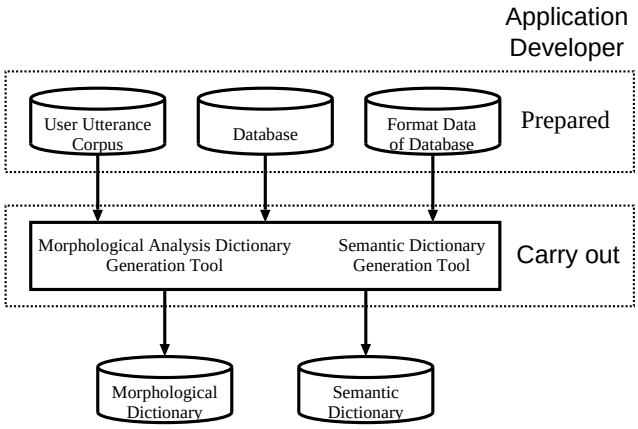

Figure 1: Task adaptation.

ble information that consists of keyword_id, reference_id and keyword_string is generated.

3.2. Generation of the dictionary for the morphological analysis

The dictionary for the morphological analysis is created from the database, user utterance corpus and field information. In this study, the nouns (4,008 words) and the keywords (5,503 words) were registered in the noun dictionary, and the authors names (5,139 words) and organization names (12 words) in the total of 3,092 papers were respectively registered in the proper noun dictionary.

3.3. Generation of the semantic dictionary for the semantics analysis

Next, the semantics dictionary is created. 495 utterances observed in retrieving information were used as the user utterance corpus. Information on the case frames of verbs and adjectives which are found in the user utterance corpus is extracted from the EDR dictionary. The case frame information of 10 adjectives and 33 verbs was extracted from the user utterance corpus.

3.4. Generation of slot information

SQL is chosen as the language for retrieving the database. From the semantic representation by the case frame of a verb, it is necessary to extract the retrieval conditions. If the retrieval conditions are expressed in the slot, the retrieval condition for "Is there a paper on speech recognition?" will consist of several constituents as shown in Table 1.

Table 1: Filled retrieval slot.

"Is there a paper on speech recognition?"

id	value
keyword	Speech Recognition
author	

3.5. Retrieval module

For information retrieval from the database, the RDBMS called PostgreSQL is used. It is possible to semi-automatically carry out the database registration to PostgreSQL using the database and the format data prepared by the developer of the system.

The algorithm which obtains the retrieval result is shown in the following.

1. SQL is input in PostgreSQL, and the paper IDs are acquired.
2. According to the number of the retrieved papers, the following procedures are carried out.
   - **none:** The prompt is presented to the user.
   - **appropriate:** From the obtained paper IDs, SQL for acquiring the paper title, the paper author, and so on is generated.
   - **too much:** The users is prompted to enter additional retrieval conditions.

3.6. Dialogue management module

The results of the retrieval and responses generated by the system are separately displayed for the user on the monitor. An example of using the system is shown in Figure 2.

When there are paper catalogues as some retrieved results, for example, in Figure 2, if the user inputs an utterance "Please display detailed information of the second paper. ", the system displays the detailed information like paper abstract of the second paper in the paper catalogue.

4. Preliminary Experiments

In order to evaluate the system, we investigated its performance through the preliminary evaluation experiments.

The domain of the system was literature retrieval from a database. 495 utterances for this domain were collected through WOZ, and 395 utterances were used to construct the word bigram and class bigram for speech recognition, and 100 utterances were used as a test-set. For comparison, 200 utterances which is included 395 utterances are selected to construct bigram too.

The transcribed 100 text utterances were uttered by three male speakers at a soundproof recording room.

Table 2 shows the results for each bigram models. In word-class and class-only bigram models, the coverage(Cov.) is 100% because all unknown words are registered with word dictionary for speech recognition. As shown in Table 2, the class-word bigram models show better results than word bigram and class bigram models.

In the Interpretation part, we extracted 37 sentences out of 100 sentences which were grammatically less complex, furthermore we rewrited them to sentences like speaking to a machine cooperatively. Using 37 modified sentences, we tested the Interpretation part. The rates of correct interpretation and correct retrieval were 73.0% and 45.9%, respectively.

```
===== System Output =====
This is a document retrieval system. Please input retrieval conditions.

===== Speech Input =====
       Is there a paper on the multi modal?  <=== input utterance
input: There is a paper on the multi modal.  <=== recognized sentence

===== System Output =====
23 papers were found.
Please input additional retrieval conditions.

===== Speech Input =====
       This is related to Internet.
input: This is related to Internet.
===== System Output =====
3 paper were found.
===== Retrieval Result =====
No.1  A. Nakashima, et al.:"Inteligent network for personal move communication",Institute
      of Electronics, Information and Communication Engineers Journal of Japan, 1995)
No.2  K. Ono, et al.:"Development of new generation communication network",Institute
      of Electronics, Information and Communication Engineers Journal of Japan, 1995)
No.3  H. Aiso, et al.:"Future prospects of information highway",Institute
      of Electronics, Information and Communication Engineers Journal of Japan, 1995)
===== System Output =====
The retrieval result is the above.
```

Figure 2: Dialogue example.

Currently, we are evaluating the Morphological Analysis part, the Interpretation part and the Dialogue management module, and improving some defects.

Table 2: Word accuracy rate(Acc.) and word correct rate(Cor.) using word bigram and class bigram.

training	bigram	Cor.[%]	Acc.[%]	Cov.[%]
dialogue 200	word only	65.4	57.0	89.5
dialogue 200	class only	70.6	66.9	100
dialogue 200	word-class	75.1	72.7	100
dialogue 395	word only	71.7	65.9	94.0
dialogue 395	word-class	77.5	68.3	100

5. Summary

The components of a spoken dialogue database retrieval system having high generality or portability were developed, and a prototype of the document retrieval system was produced. The semantics analysis, speech recognition modules and the language modules were completed and response generation module and dialogue management module were under development.

In the semantics analysis module, since the system has high generality in terms of the the database retrieval task, the generality of this system can be increased by assigning task-dependent information only to the dictionary. The speech recognition module is task-independent.

In the near future, the generality of the language model module, dialogue management module and response generation module will be improved and integrated.

REFERENCES

1. A. Kai and S. Nakagawa. Investigation on unknown word processing and strategies for spontaneous speech understanding. In *Proc. of EUROSPEECH '95*, pages 2095–2098, 1995.
2. T. Itoh, M. Hidano, M. Yamamoto, and S. Nakagawa. Spontaneous speech understanding for a robust dialogue system. In *Proc. of NLPRS '95*, volume 2, pages 538–543, 1995.
3. A. Denda, T. Itoh, and S. Nakagawa. A robust dialogue system with spontaneous speech and touch screen. In *Proc. of ICMI '96*, pages 144–151, 1996.
4. S. Kogure, T. Itoh, and S. Nakagawa. A semantic interpreter for a robust spoken dialogue system. In *Proc. ICMI '99*, pages II 61–66, 1999.
5. S. Nakagawa, S. Kogure, and T. Itoh. A semantic interpreter and a cooperative response generator for a robust spoken dialogue system. *IJPRAI*, 14(5), 2000.
6. S. Kaspar and A. Hoffmannn. Semi-automated incremental prototyping of spoken dialog systems. In *Proc. of ICSLP '98*, pages 859–862, 1998.
7. T. Brondsted, B. Bai, and J. Olsen. The reward service creation environment. an overview. In *Proc. of ICSLP '98*, pages 1175–1178, 1998.
8. S. Sutton, R. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y.Yan, E. Kaiser, B. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, D. Massaro, and M. Cohen. Universal speech tools:the cslu toolkit. In *Proc. of ICSLP '98*, pages 3221–3224, 1998.
9. M. Sasajima, T. Yano, and Y. Kono. Europa: generic framework for developing spoken dialogue systes. In *Proc. of EUROSPEECH '99*, pages 1163–1166, 1999.
10. S. Kurohashi, T. Nakamura, Y. Matsumoto, and M. Nagao. Improvements of japanese morphological analyzer juman. In *Proc. of International Workshop on Sharable Natural Language Resources, Nara*, 1994.
11. A. Kai, Y. Hirose, and S. Nakagawa. Dealing with out-of-vocabulary words and speech disfluencies in an n-gram based speech understanding system. In *Proc. of ICSLP 98*, pages 2427–2430, 1998.
12. S. Nakagawa. Architecture and evaluation for spoken dialogue systems. In *Proc. of International Symposium on Spoken Dialogue (ISSD '98)*, pages 1–8, 1998.