A TELEPHONE NUMBER INQUIRY SYSTEM WITH DIALOG STRUCTURE

Hsien-Chang Wang and Jhing-Fa Wang

Institute of Information Engineering, National Cheng-Kung University
1, Ta-Hsueh Road, Tainan City, Taiwan, R.O.C.
{wangjf, wangsj}@server2.iie.ncku.edu.tw

ABSTRACT

A Telephone Number Inquiry System (TNIS) answers caller the phone number he/she wants to know. Traditional system requires the caller to know the full name of the party [1] [7]. If the caller forgets the name, the system fails to retrieve correct information for the caller.

In this paper, we propose a novel TNIS with dialog structure that can let caller use a more flexible method while inquiring, i.e., the caller may interact with our system to inquire the phone number by providing just the working, researching area, the surname, or the title, etc. Our system takes the telephone speech as input, after generating the word sequence, it performs a maximum likelihood key-feature matching with knowledge base. If necessary information is not derived, interactive dialog manager is activated to resolve the caller's requirement. The experimental results show that our novel approach can make the system more natural.

1. INTRODUCTION

In this decade, dialog systems have been broadly researched. The goal is to let a machine properly interacts with the user to provide service for the user. Applications such as railway ticket reservation[5], automobile purchasing guide [4], and restaurant information tutoring system [6] have been presented to the public and demonstrate the ability of serving people.

Our system is defined as a telephone number inquiry system (TNIS) which can interact with the caller using spoken language and finally provides the telephone number for the caller. From the observation of the real dialog between the operator and the caller, we classify the inquiry into two types. One is *Clear inquiry* which means the caller provides enough information for database access. The other type is *fuzzy inquiry* which means the caller provides only partial information. Traditionally, a TNIS can deal with only clear inquiry [1] [7]. In this paper, we propose a novel system which takes into account the working/researching area information to handle fuzzy inquiry problems. Two different inquiry examples are shown below.

- **Clear inquiry**: "Please tell me the phone number of Chang San in the department of computer science and information engineering".
- **Fuzzy inquiry**: "I want to know the phone number of the professor who majors in speech coding".

The architecture of our system is shown in Figure 1. Our system is designed modularly, it contains six modules which are briefly described below.

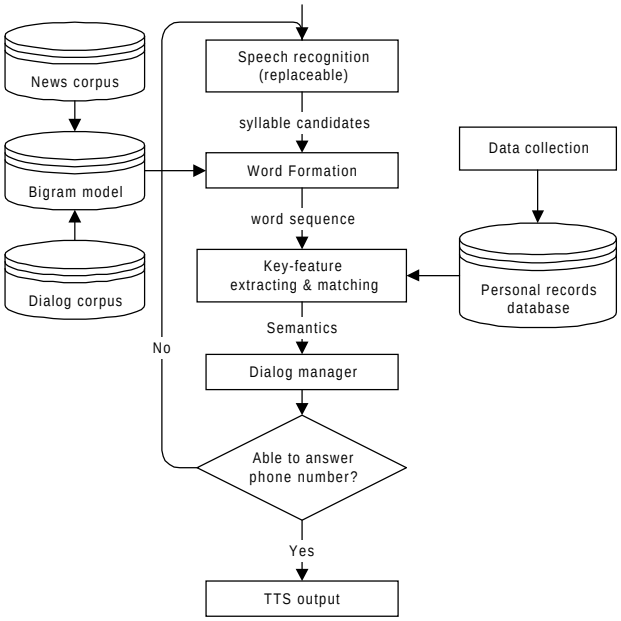

Figure 1. The system architecture of the telephone number inquiry system.

- **Language modeling:** building statistical language model needed for speech recognition.
- **Telephone speech recognition:** transfer input speech into syllable candidates.
- **Word formation:** transfers the syllable candidates into word sequence.
- **Key-feature matching:** matching the features in the input with those in the personal database and generate the semantic frame.
- **Dialog manager:** decide how and what to respond.
- **Text-to-speech output:** output speech to the caller.

These modules are written as Microsoft dynamic link libraries(DLL). Each module can be tested, updated or replaced independently. This makes our system more testable and configuriable. Besides, our system is aimed to be practically used in many companies or institutes for office telephone-transfer automation. In order to port our system from one company (institute) to another company (institute) easily, we collect many dialog structures from different companies to set up a dialog-structure database for the common use. Armed with this special feature, a new company only needs to provide the personnel database and need not provide dialog corpus to port

the system.

We introduce these six components in the following sections. The experiments and results are described in Section 5. Finally, we give the conclusion in Section 6.

2. SPEECH RECOGNITION & WORD FORMATION

2.1. Telephone Speech Recognition

We use a speaker independent continuous speech recognition system as the front-end to translate the input speech into syllable candidates. It is basically an HMM model with 3 consonant-states and 5 vowel-states. The features of each frame are 12-th order MFCC cepstrum plus delta and delta-delta MFCC cepstrum coefficients. The training speech is derived from MAT telephone speech database [9]. For each syllable of the input speech, there are five candidates which form the syllable lattice. The syllable lattice of the input speech is highly ambiguous, so it needs to be decoded into word sequence. We use a statistical language model approach to find the most likely word sequence for further processing.

2.2. Corpus and Language Model

There are two word bigram language models to be construct in this task. The first one is the general language model. It is trained using news corpus which is collected from the newspaper and WWW. The corpus contains about 10 million of Chinese words.

The other is the domain specific language model which is trained using the telephone dialog corpora. To collect the corpus for building language model, the real dialogs between the caller and the operator are recorded by setting up a recording device in the operating room. In this way, we collect about 500dialogs. Following are two examples of dialog between the caller and the operator.

***Operator**: This is Cheng-Kung Univ., how may I help you?*
***Caller**: I want to know the phone number of professor Chang.*
***Operator**: In which department?*
***Caller**: Oh, he is in the Department of Computer Science and Information Engineering.*
***Operator**: There are two professors in this department, Chang Yi and Chang San, which one do you want?*
***Caller**: Chang San, please.*
***Operator**: Professor Chang San, his phone number is 62111.*

***Operator**: This is Cheng-Kung Univ., how may I help you?*
***Caller**: Would you please find the phone number of professor Huang in the Institute of Information Engineering?*
***Operator**: Professor Wang? What is his first name?*
***Caller**: No, no, I want to call professor Huang Zong-Li.*
***Operator**: Professor Huang Zong-Li, please dial 62222.*

The language models are built using the simplified expression for the bigram case as shown in Formula 1.

$$P(W_0,...W_n) = \prod_{i=1}^{n} p(W_i|W_{i-1}) \quad (1)$$

Where $(W_0,\ldots,W_n)$ is the word sequence, $P(W_i|W_{i-1})$ is the probability of word-pair (W_{i-1}, W_{i-1}) appears together. We use the Viterbi algorithm to find the word sequence which has maximum probability [8].

The combined language model is generated by using an interpolated method. The word error rate is reduced by about 20% when we use the combined language model instead of the general one.

3. KEY-FEATURE & SEMANTICS

3.1. Representation of Personal Records

To form the semantic frame, the word sequence generated by the word formation unit is then matched with the personal record database using the key-features. Each personal record in our system contains seven **key-features**, 1) first name, 2) surname, 3) title, 4) sex, 5) department, 6) working/researching area, and 7) phone number.

3.2. Key-feature Matching

Among all of the personal records, those which have most likely features when matching with the input are chosen as the candidates. We propose the key-feature matching approach which takes into account the synonyms of the input. For example, if the input contains a string "teacher Chang", then the personal record which has surname "Chang" and title "professor" or "associate" will match the input. Some synonyms defined in our system are listed below. Note that *professor* and *associate professor* are synonyms since people usually confused about them.

- *Title:* teacher, professor, associate professor,…
- *Sex:* female, Ms., Mrs., lady,…
- *Researching Area:* speech coding, speech compression, audio compression,…
- *Working Area:* management, administration,

With this synonymy defined, our system is able to deal with more natural inquiry input.

3.3. Semantic Frame

Every input sentence has a corresponding speech act type (SAT) which represents the meaning or the intention of the input [3]. In order to determine what action should be performed after accepting the input, the result of the key-feature matching and the speech act type have to be decided first. We use the semantic frame to keep such information. A semantic frame contains a speech act type plus seven semantic slots corresponding to the seven key-features in the personal record. The matched features of that record are filled into the semantic slots.

To determine the SAT of the input, we use the same technique as proposed in our previous work [2], i.e. check if the input sentence contains certain pattern of SATs. We define six SATs

in this application domain, 1) inquiry for phone number, 2) refinement, 3) repair, 4) confirm, 5) greeting, 6) undecidable.

4. DIALOG MANAGER

Because of the ambiguity of the speech recognition and the multiple results of database access, it is necessary for the system to interact with the caller to derive the phone number. The block diagram of the dialog manager is shown as in Figure 2. The grayed round block means the SAT of the caller, the white block represents the corresponding action and response of the system which is described below.

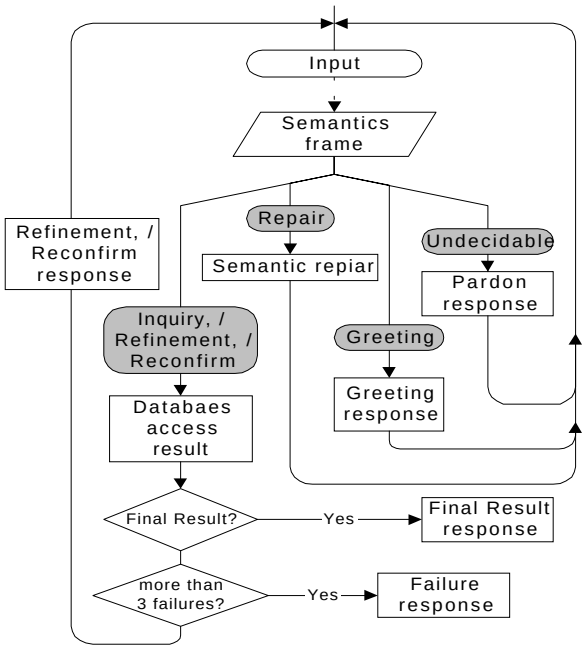

Figure 2. The interactive dialog manager

4.1. Dialog Strategies

In order to make the dialog smoothly, we have some strategies when interacting with the caller.

- **Brief strategy:** We make the response sentence as brief as possible. Since the response is output by a text-to-speech system, it will be unsatisfied if the "machine-like" sound of the system is too long. For example, if the matched results are either "professor Chang in the Institute of Information Engineering (IIE)" or "professor Wang in the IIE", the response will be merged as "professor Chang and Wang in the IIE" in stead of just listing the details of two matched records.
- **Mixed initiate strategy:** Another strategy is that we use a mixed initiative dialog approach. The system usually prompts the caller for information needed for database query when there remain some semantic slots unfilled. In addition, the caller is allowed to ask any question spontaneously.
- **Reconfirm and Repair strategy:** To make sure the results of the system are all correct, we ask the caller to reconfirm them in the final turn of the dialog. If caller finds that the system does not recognize his intention correctly, he may repair it at the next turn of the dialog.
- **Simple strategy:** To simplify the dialog, we allow the caller inquire one phone number at a time. Also, we do not ask the user to answer more than two questions in each turn.

4.2. Dialog Response Generation

The interaction of the dialog depends on the semantics derived in Section 3.3. and some rules we define. In different situation, proper rule is applied to generate the relative responding sentence. The generations of the responding sentence are classified below.

- ***Greeting response:*** In the beginning and ending of a dialog, the greeting response such as *"How may I help you?"*, *"Good bye"* are responded.
- ***Refinement response:*** If there are more than 3 matched records, our system requires the caller to input more information. For example,

  Caller: *"What is the phone number of the professor in the Institute of Information Engineering?"*

  System: *"What is the researching area or the name of this professor?"*
- ***Reconfirming response:*** When the system has derived the phone numbers of the inquiry, it asks the caller to confirm it, such as *"Professor Wang Jhing-Fa, Phone number 62111, is that what you want?"*
- ***Repair response:*** If the caller corrects the response of the system, the old semantic slots are modified and the key-features newly derived is filled into the semantic frame. The dialog continues with the new semantics.
- ***Result response:*** When the phone number is determined, our system generates the result to the caller.
- ***Pardon response:*** If the system could not make a decision for the recognized word sequence or the information that caller provides conflict with the database records, it will prompt to the caller and ask him to input again.
- ***Failure response:*** If there are more than three times that the system cannot properly proceed with the input, we claim that it is a failure dialog, and the caller is switched to the human operator.

5. EXPERIMENTS AND RESULTS

Our experimental environment consists of a Pentium 200 PC with a 4-port Dialogic card. The personal database contains about 200 records and the response time for each input is about one second. The experiment is carried out with 100 callers (65 males and 35 females in our school.) They are asked to make more than one phone call during the experiment. In this way, we have 165 dialogs for the testing of our system. The average age of the caller is 24 years old. Most callers are cooperative, that means they usually obey the response of the system.

The experiments contain two parts, the semantics accuracy and the dialog accuracy. With the aid of interactive dialog manager, although the speech recognition correct rate is not high enough,

the performance of this telephone number inquiry system is not bad. Most testers represent that they are satisfied with the result of our research.

5.1. Experiment on Semantics Accuracy

The speech recognition accuracy is tested for each semantic slot and speech act type. The results are shown in the upper part of Table 1. The average correct rate of the speech recognition over semantic slots is 77.3 %. Note that, since the Chinese surname is highly confused, it causes most of the speech recognition errors.

For each input, if the semantics slots and the speech act type are correctly filled into the semantics frame, then it is an accurate processing of semantics. There are about 650 input sentences in the 165 dialogs. Among these 650 input sentences, 478 of their semantics can be determined correctly. The accuracy of semantics is 73.5% in this experiment as shown in the bottom row of Table 1.

Type	Correct rate
First name	83.7 %
Surname	66.4 %
Title	80.5 %
Sex	70.2 %
Department	78.3 %
Working/Researching area	74.1 %
Speech act type	87.9 %
Average of semantic slots	77.3 %
Semantics Frame	73.5 %

Table 1. Recognition rate of semantic slots and semantics frame

5.2. Experiment on Dialog Accuracy

If the caller can finally know the desired phone number from our system within three repairs, then it is a successful dialog. Among 165 testing dialogs, 113 of them are clear inquiries and 52 of them are fuzzy. 85 dialogs are successful in clear inquiries and 28 are failed. For fuzzy inquiries, 39 of them are successful and 11 of them are failed. The total dialog success rate is about 75.1%. The dialogs are completed in an average of four turns. The result of the dialog accuracy is shown in Table 2.

	clear inquiries	fuzzy inquiries
# of dialogs	113	52
# of successful dialog	85	39
successful rate	75.2 %	75.0 %

Table 2. Test results of the dialogs accuracy

6. CONCLUSIONS AND FUTURE RESEARCH

We have presented a more natural and flexible telephone number inquiry system (TNIS) with dialog structure. This system is capable of dealing with telephone number inquiry by using either full-name search or the related information search such as title, researching/working area, etc. In addition, our system is designed modularly and embedded with a general common dialog structure. These will make our system more testable, configuriable and portable.

Although there remain some problems to be solved such as robust speech recognition over the telephone network and out of vocabulary problems, the experimental results show that the interactive dialog manager we proposed achieves the goal to be an user friendly system.

Our future research will aim to improve our system accuracy, such as robust speech recognition, repair policy. And, we will study how to rapidly transfer our system from telephone number inquiry system into another practical system —hospital register automating system.

7. REFERENCES

[1] Hermann H. and Alex W. *Recognition of Spelled Names Over the Telephone*, ICSLP'96 Vol. 1.

[2] Hsien-Chang Wang, Jhing-Fa Wang, and Yi-Nan Liu, *A Conversational Agent for Food_ordering Dialog Based on VenusDictate*, Proceedings of ROCLING X International Conference 1997, pp.325-334.

[3] Y. J. Yang, L. F. Chien and L. S. Lee, *Speaker Intention Modeling for Large Vocabulary Mandarin Spoken Dialogues*. ICSLP'96, Vol. 2.

[4] Helen M., Senis B, and Victor Zue, et al. *WHEELS: A Conversational System in the Automobile Classification Domain*, ICSLP'96 Vol. 1.

[5] S. Bennacef and L. Lamel et al., *Dialog in the RAILTEL Telephone-Based System*, ICSLP'96 Vol. 1.

[6] H. Tsuboi and Y. Takebayashi, "A real-time task-oriented speech understanding system using keyword spotting," Proc. ICASSP, pp.197-200,1992.

[7] Frank Seide and Andreas Kellner, *Toward an Automated Directory Information System*, EuroSpeech'97 Vol. pp.1327-1330.

[8] S. Furui and M. M. Sondhi, *Advances in Speech Signal Processing*, Marcel Dekker, Inc., pp.652-699, 1992.

[9] Hsia-Chuan Wang, *MAT – A Project to Collect Mandarin Speech Data through Telephone Networks in Taiwan.* International Journal of Computational Linguistic Chinese Language Processing, Vol. 2. No. 1, pp.73-89, 1997.