A SPOKEN DIALOG CORPUS FOR CAR TELEMATICS SERVICES

Masahiko Tateishi and Ichiro Akahori
Research Laboratories, DENSO CORPORATION
Nisshin city, Aichi, 470-0111, JAPAN, mtatei@rlab.denso.co.jp, iakahori@its.denso.co.jp

Scott Judy, Yasunari Obuchi, Teruko Mitamura, and Eric Nyberg
Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA 15213, {scottj, obuchi, teruko, ehn}@cs.cmu.edu

ABSTRACT

Spoken corpora provide a critical resource for research, development and evaluation of spoken dialogue systems. This paper describes a spoken dialog corpus used in the design of CAMMIA (Conversational Agent for Multimedia Mobile Information Access), which employs a novel dialog management system that allows users to switch dialog tasks flexibly. The corpus for car telematics services was collected from 137 male and 113 female speakers. The age distribution of speakers was balanced in the five age brackets of 20's, 30's, 40's, 50's, and 60's. Analysis of the dialogs gathered reveals that the average number of dialog tasks per speaker is 5.0. The most frequent tasks in the corpus focus on traffic information, such as route guidance and traffic congestion, followed by restaurant guidance and sightseeing information. It is also shown that to cover 90% of the tasks in the corpus, the dialog system must handle at most six most frequent tasks. Analysis of speaker utterances shows that the vocabulary size is approximately 5,000 words. These results are used to develop and evaluate ASR as well as dialog scenarios.

1. INTRODUCTION

Telematics is the emerging industry of communication, information, and entertainment services delivered to motor vehicles via wireless technology. The telematics system must provide HMI that allows the drivers to operate the device, system or service easily and without any risk to traffic safety. Spoken dialog system is considered to be a suitable HMI for telematics by allowing the driver to keep "hands on the wheel, eyes on the road".

The Conversational Agent for Multimedia Mobile Information Access provides a framework of client-server type of spoken dialog systems in mobile, hands free environments[1]. The goal of CAMMIA is to realize large-scale speech dialog systems that can handle a variety of information retrieval tasks. CAMMIA is based on VoiceXML, which is proposed as a standard markup language by W3C. The client is an in-vehicle terminal with an ASR, VoiceXML interpreter, and TTS, whereas the server is a computer with a dialog manager (DM). The client recognizes the driver's utterances according to the VoiceXML dialog scenarios and transmits the recognition result as the query to the server. The server searches its database and transforms the search results into VoiceXML dialog scenarios.

The novel aspect of CAMMIA is the natural conversational interaction between the user and the system including the DM that allows the user to change dialog tasks flexibly. Many of the issues to support natural spoken dialog can be cast in terms of the modeling of human behavior observed in large collections of spoken or written data. Specifically, the modeling includes defining the lexicon and grammar of ASR as well as designing suitable dialog scenarios used by the DM.

Human-computer dialog differs from human-human dialog in various aspects including linguistic complexity[2]. However, examination of human-human dialogs is a natural first step in the process of modeling human dialog behavior [3]. The modeling approach requires very large quantities of task oriented linguistic data. To meet this requirement we collected a spoken dialog corpus for car telematics services. Section 2 outlines the system architecture of CAMMIA. Section 3 explains the spoken corpus collection. Section 4 describes the analysis of the corpus, followed by conclusions and future work.

2. SYSTEM ARCHITECTURE OF CAMMIA

Figure 1 depicts the current system architecture of CAMMIA. It consists of in-vehicle terminals and the server connected by wireless network. The client consists of an ASR, a VoiceXML interpreter, and a TTS, whereas the server consists of DM, database, and dialog scenario.

The spoken dialog corpus is used to evaluate the lexicon and grammar of ASR and the suitable dialog scenarios that represent particular dialog tasks such as traffic information. It is also used to evaluate the coverage of dialog tasks. We will discuss those evaluations in Section 4.

3. SPOKEN CORPUS COLLECTION

In this section, we describe how the spoken corpus was collected and transcribed.

3.1 The Speakers

Not only the lexicon and grammar of speaker's utterances but also the kinds of dialog task could significantly differ by the gender and generation of the speaker. Therefore, it is desirable to balance the gender and age range of the speakers. We collected the spoken corpus from 250 speakers consisting of 137 male and 113 female speakers. The age distribution of speakers was also balanced in the five age brackets of 20's, 30's, 40's, 50's, and 60's. They were residents of Tokyo Metropolitan Area and 235 of them had driver's licenses. Fifty of them had experience to use navigation systems

3.2 The Experimental Setup

The spoken corpus was recorded in a studio in Tokyo. The studio consisted of two rooms. The speaker was in one room and an operator who gave driving information was in the other room. The speaker's room corresponded to a car equipped with the in-vehicle terminal, whereas the operator's room corresponded to the server. The speaker and the operator could talk each other using microphones and headsets but the two rooms were completely separated so that no nonverbal interactions occurred between them. The operator answered the query from the speaker as a travel agent who had the following information:

- Real-time traffic information
- Restaurant information
- Sightseeing information
- Hotel and ryokan (Japanese inn) information

In order to handle task oriented dialogs that include tourist information skillfully, the operator must have an expertise in travel information in addition to the capability to speak correct Japanese. In order to meet these requirements we employed an experienced female announcer who had a job experience in travel agency.

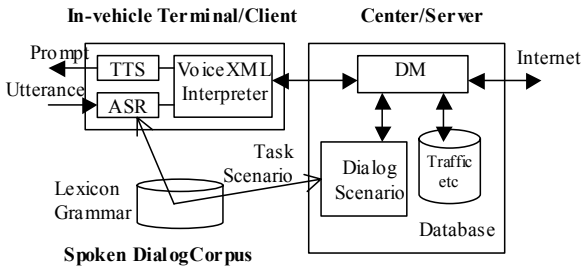


Figure 1: System Architecture of CAMMIA

3.3 The Task and Instructions

The speaker was told to pretend that he/she was in a car equipped with the in-vehicle terminal and was going to visit one of the two sightseeing areas, i.e., Hakone and Izu, from Tokyo. Both the sightseeing areas are very popular and well known to residents in Tokyo Metropolitan Area. The speaker was instructed to talk with the operator in order to obtain driving information. A preliminary survey on 302 drivers showed the three most needed driving information were traffic, restaurant, and sightseeing information. Therefore, we first tried following three predefined tasks:

Task1. Obtain route guidance to the destination.
Task2. Find a restaurant at the destination for lunch.
Task3. Find tourist attractions to visit after lunch.

After collecting the spoken dialog corpus from the first 50 speakers, we found predefined tasks had several problems. Predefined tasks failed to make speakers pretend they were on a travel enough to generate questions relevant to the tasks, indicating the speakers were not well motivated. So the operator sometimes had to halt the conversation and instruct the speakers what kind of questions they could ask. It was also found that predefined tasks might limit the types of dialog tasks.

To cope with these problems we divided the remaining 200 speakers into 40 groups each of which consists of five speakers. Each group was instructed to choose Hakone or Izu as the destination and to discuss a driving plan of overnight trip according to their interests. After the discussion each speaker generated two sets of dialog tasks *A* and *B* relevant to the driving plan. The set *A* and set *B* consisted of tasks to obtain information needed before starting the trip on the first day and before leaving the hotel/ryokan on the second day, respectively. The recording of the dialog was also divided in two sessions *A* and *B* that correspond to the first day and the second day, respectively. Then we instructed the speaker to do the task sets *A* and *B* at the corresponding sessions, respectively.

L:はいドライブ情報センターです

Driving information center, may I help you?

R:すいません

Well..,

L:はい

Yes

R:えー{R東京駅}から

Eh, From {R Tokyo Station} …

L:はい

Yes

R:えー{P箱根}までの行き方を教えていただきたいんですが

eh, to {P Hakone}. Please give me the route guidance.

L:はいえ{R東京駅}からえー{P箱根}ですねー

From {R Tokyo Station} to {P Hakone}.

R:はい

Yes.

Figure 2: Example of Transcribed Text

Letting the speakers to plan the trip significantly improved motivating them. The first 50 speakers needed 68 instructions by the operator, whereas the last 200 speakers needed only one instruction.

3.4 Transcriptions and Annotation

Each dialogue was transcribed into text by hand from the audiotapes. The dialogs were segmented into turns. Lines with prefix 'L:' and 'R:' in the text files represent the operator's and speaker's turn, respectively.

Next, we annotated proper nouns in the corpus in brackets of the format {*X* name}, where *X* is one of the following 5 letters: **A**, **P**, **R**, **S**, and **W** that represent names of tourist attraction, place, railway facility, shop/restaurants, and traffic facility, respectively. Traffic facility includes names of road, entrance/exit of expressway, etc. Figure 2 shows an example of transcribed text after annotation. English translations appear below the utterances. The dialog starts with the operator's utterance "Driving information center, may I help you?" The speaker wants the route guidance from Tokyo Station to Hakone.

The annoted text will be segmented into morphemes using ChaSen[4] – a Japanese morphological analyser.

Annotating proper nouns is useful not only for correct morphological segmentation but also for designing the lexicon and grammar of ASR. It is also convenient for designing class *N*-grams used by the ASR, where proper nouns with the label *X* form a class *X*.

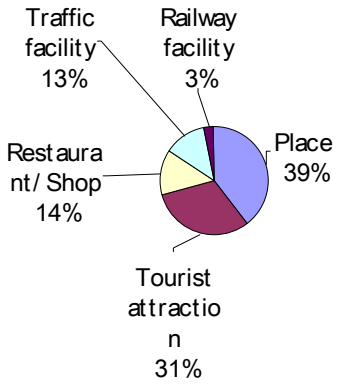


Figure 3: Ratio of vocabulary size of proper noun

4. ANALYSIS OF THE SPOKEN CORPUS

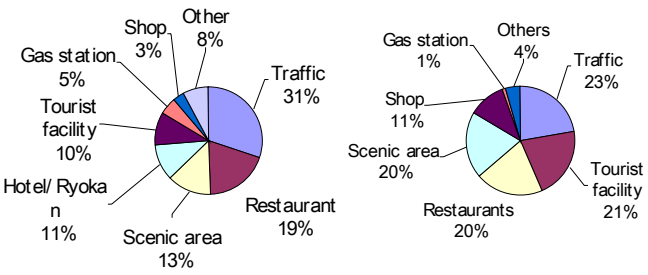
Our spoken dialog corpus plays several roles in the design of CAMMIA, e.g.: a) as a reference for creating the ASR grammar, b) as a test suite for testing the ASR grammar, c) as a source for the real lexicon, d) as a guide for identifying the highest frequency sentences and words, and e) as a reference for possible scenarios in the conversation. In this paper we focus on the statistics of the speakers' utterances to determine the lexicon of ASR. We also discuss the types of the individual tasks in the corpus in terms of the coverage.

4.1 Statistics of Speakers' Utterances

The spoken corpus consisted of 450 conversations comprising of 33,072 speaker turns. The speakers' utterances were extracted and analyzed to determine the lexicon of ASR.

The vocabulary size of speakers' utterances was 4,603 consisting of 999 proper nouns and 3604 words other than proper nouns. The former set of proper nouns varies according to the sightseeing areas to be guided. On the others hand the latter set would be commonly used in similar tasks. From this observation, the lexicon size of ASR is approximately 4,000 to 5,000 to recognize the speaker's utterances when applied to the same kind of tasks.

Figure 3 illustrates the ratio of annotated proper noun. Four types of proper noun, i.e., place, tourist attraction, restaurant/shop, traffic facility, summed up to 97%. The first two types summed up to no less than 70%. Although the ratio depends on the sightseeing areas to be guided and the operator's dialog strategy, this suggests the ASR should support more names of place and tourist attraction than the others.



Task *A* Task *B*

Figure 4: Ratio of task frequency in the set *A* and *B*

4.2 Statistics of Tasks

Analysis of the dialogs gathered reveals that the total number of dialog tasks of 250 speakers is 1253, indicating the average number of tasks per speaker is 5.0. The most frequent tasks in the corpus focus on traffic information (28%), followed by restaurant guidance (21%) and sightseeing information, i.e., tourist facility (15%) and scenic area (15%).

As described in Section 3.3, we let 200 speakers to generate two sets of dialog tasks *A* and *B* according to their interests. Those two sets corresponded the task before starting the trip on the first day and the task before leaving the hotel/ryokan on the second day, respectively. Total number of task frequency in the set *A* and *B* are 671 and 429, indicating the average numbers of tasks per speaker are 3.4 and 2.1, respectively.

Figure 4 shows the comparison of task frequency in the set *A* and *B*. On the second day hotel/ryokan information disappeared and sightseeing information, i.e., tourist facility and scenic area, attracted more interests. It is also worthy to note that the task frequency of shop increased from 3% to 11%. This is because many speakers were interested in purchasing souvenirs such as dried fish on their way home. Although the spoken corpus was collected in a studio, these observations prove the speakers' travel experiences were successfully reflected in the task selection, indicating the effectiveness of the task setting and the instructions we employed in Section 3.

5. CONCLUSIONS AND FUTURE WORK

A spoken corpus for car telematics services was collected from 137 male and 113 female speakers. Analysis of the spoken corpus revealed the vocabulary size of speakers' utterance was 4,603 consisting of 999 proper nouns and 3604 words other than proper nouns. To cover 90% of the tasks in the corpus, the dialog system must handle seven tasks. These results are used to develop and evaluate ASR as well as dialog scenarios of CAMMIA.

The spoken corpus has several issues to be improved for developing ASR grammar and the dialog scenario:

(i) The operator does not talk like a computer.
- The operator uses ambiguous expressions, such as "the route is congested a little bit heavily".
- The operator does not always state things in a succinct way.

(ii) The speaker does not act like he is talking to a computer
- There are too many hesitation words such as "eto" in the dialog.

Currently these parts are corrected by hand. Apparently these issues result from human-human dialogs. To cope these problems we are planning to use the first version of dialog system based on human-human spoken dialog corpus to collect human-machine dialogs and improve the dialog system in incremental steps.

ACKNOWLEDGEMENTS

The authors thank Dr. N. Hataoka and Mr. J. Watanabe of Hitachi Ltd for their valuable support. The authors are also grateful to Dr. M. Araki, Assistant Professor of Kyoto Institute of Technology, for his staunch support of developing spoken corpus.

REFERENCES

[1] E. Nyberg, T. Mitamura, P. Placeway, M. Duggam, and N. Hataoka, "DialogXML: Extending Voice-XML for Dynamic Dialog Management", Proc. of HLT-2002, 2002

[2] C. Doran, J. Aberdeen, L. Damianos, and L. Hirshman, "Comparing Several Aspects of Human-Computer and Human-Human Dialogues", Proc. of Second SIGdial Workshop on Discourse and Dialog, Aalborg

[3] C. Cieri, "Resources for Robust Analyses of Natural Language", Conference ROMAND 2000, Lausanne, 2000

[4] NARA INSTIUTE of SCIENCE and TECHNOLOGY, "Morphological Analyzer ChaSen", http://chasen.aist-nara.ac.jp/