TOWARDS ROBUST SPOKEN DIALOGUE SYSTEM USING LARGE-SCALE IN-CAR SPEECH CORPUS

Yukiko YAMAGUCHI†, *Keita HAYASHI*†, *Takahiro ONO*†, *Shingo KATO*†, *Yuki IRIE*†,
Tomohiro OHNO†, *Hiroya MURAO*‡, *Shigeki MATSUBARA*†,
Nobuo KAWAGUCHI†, *Kazuya TAKEDA*†,

†Nagoya University Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
‡SANYO Electric Co., Ltd. 1-18-13 Hashiridani, Hirakata-shi, Osaka, 573-8534, Japan
yamaguchi@itc.nagoya-u.ac.jp

ABSTRACT

We have been studying various topics by using a large-scale corpus, which was built at CIAIR, to construct a robust and practical spoken dialogue system. The CIAIR project has developed a data collection vehicle and collected about 179 hours of multi-modal data in total.

We have transcribed the speech data by about 800 subjects, and annotated speech intentions, dependency structures, dialogue structures to the text data. We are continuing various research using the annotated data, such as speech intention understanding and speaker's knowledge acquisition. In this paper, we introduce our research activities, and present the various fruits of the in-car speech corpus.

1. INTRODUCTION

With the recent advances in continuous speech recognition technology, a considerable number of studies have been conducted on spoken dialogue systems. A lot of large-scale corpora are collected [1], and the development of the corpus-based system is also active. However, in order to use the corpus for system development it is preferable that the corpus has additional language information besides the speech data and the transcribed text.

To improve the usefulness of the corpus, we have annotated speech intentions, dependency structures, dialogue structures to the CIAIR corpus. Using the annotated corpus we have been researching speech intention understanding and information extraction, and developing practical spoken dialogue systems.

This paper introduces our research activities. Section 2 describes the CIAIR corpus. Section 3 presents the elaboration of the corpus. Section 4 describes the analysis and utilization of the corpus. Section 5 introduces the development of the spoken dialogue system which is an application of the corpus.

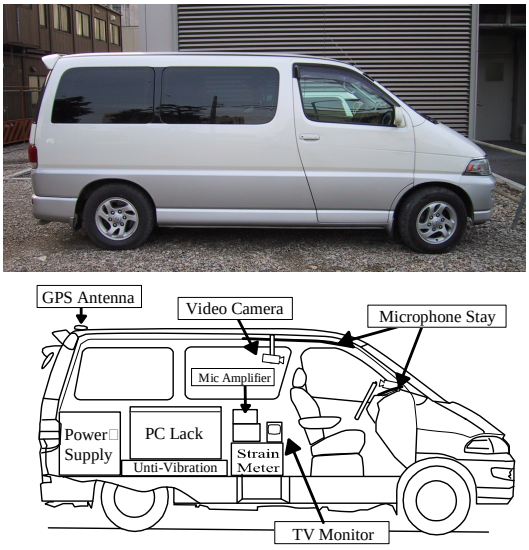


Fig. 1. Data Collection Vehicle

2. CIAIR IN-CAR SPEECH CORPUS

The Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University has collected a large-scale corpus of the in-car speech [2, 3, 4]. During the project, CIAIR members have developed a Data Collection Vehicle (DCV), which is shown in Figure 1, and collected a total of about 400 GB of data by recording the spoken dialogues by 812 drivers. While the drivers were actually driving the DCV, they made three sessions of dialogues with a human operator, with a WOZ system and with a spoken dialogue system.

The collected speech data has been transcribed into ASCII text files by hand in accordance with the rule of the corpus of spoken Japanese (CSJ) [1]. An example of a transcript is shown in Figure 2. Each utterance is divided into utter-

0028 - 02:21:353-02:24:909 F:D:I:I:

(F えっと)ラーメンが	&	(F エット)ラーメンガ	(Well,) chinese noodle
ちょっと	&	チョット	
食べたいんだけど	&	タベタインダケド	want to eat
どっか	&	ドッカ	any restaurant
ないかしら<SB>	&	ナイカシラ<SB>	there are

0029 - 02:26:691-02:32:162 F:O:I:I:

はい	&	ハイ	Yes,
この	&	コノ	
近くですと	&	チカクデスト	near here
該当する	&	ガイトースル	
お店が	&	オミセガ	restaurants
三軒	&	サンケン	three
ございます<SB>	&	ゴザイマス<SB>	there exist

Fig. 2. Example of transcription

Table 1. Size of the annotated spoken dialogue corpus

(a) Speech intention annotated corpus

number of dialogues	3,641
number of utterance units	35,421
Driver's utterances	16,224
Operator's utterances	19,187
number of LIT types	95

(b) Dependency structure annotated corpus

number of utterance units	13,756
number of dependency relations	45,053

(c) Dialogue structure annotated corpus

number of dialogues	789
number of utterance units	8,150
number of LIT sequence types	657
number of structural rules	297

ance units by a pause of 200 msec or more. The transcribed text has the tags of the grammatically ill-formed linguistic phenomena such as fillers, hesitations and so on.

We have investigated the feature of the CIAIR corpus [5]. For the drivers' utterances the number per utterance unit of fillers, hesitations and slips, is 0.31, 0.06, 0,03, respectively. And we have realized that the drivers' utterance are affected by driving conditions, accelerator and brake operation, and steering-wheel operation.

In these research which are introduced in this paper, we use the restaurant guide dialogues between drivers and operators and between drivers and the WOZ system.

3. ELABORATION OF IN-CAR SPEECH CORPUS

To advance the construction of robust spoken dialogue systems, we have annotated additional information such as the speech intention, dependency structure, dialogue structure. Table 1 shows the size of the annotated spoken dialogue corpus.

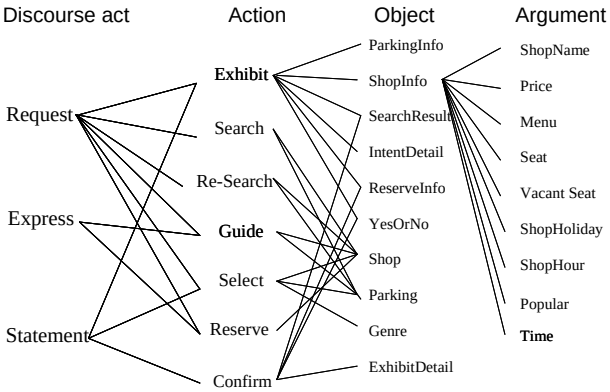


Fig. 3. Layered intention tag (a part of)

#	Speaker	Utterance	LIT
0028	D	(えっと)ラーメンがちょっと食べたいんだけどどっかないかしら I want to eat Chinese noodle. Are there any restaurants?	Request + Search + Shop
0029	O	はいこの近くですと該当するお店が三軒ございます Yes, There are three restaurant near here.	Statement + Exhibit + SearchResult + ShopName
0030	O	ラーメンギョーザの金龍ラーメンさっぽろ亭春帆亭でございます Kinryu-Ramen of Chinese noodles and dumplings, Sapporo-tei, Shunban-tei.	Statement + Exhibit + SearchResult + ShopName
0031	D	さっぽろ亭でお願いします I'd like to go to Sapporo-tei.	Statement + Select + Shop
0032	O	はいさっぽろ亭ですと駐車場がございません Sapporo-tei has no parking lot.	Statement + Exhibit + ShopInfo + Parking
0033	O	よろしかったでしょうか Is is all right?	Request + Exhibit + YesorNo
0034	D	じゃ駐車場があるところお願いします Well, please let me know the restaurant with a parking lot.	Request + Re-search + Shop
0035	O	はい金龍ラーメンと春帆亭ですと駐車場がございますが Kinryu-Ramen and Shunban-tei have parking lots.	Statement + Exhibit + SearchResult + ShopName
0036	D	じゃ金龍ラーメンお願いします Well, I'll go to Kinryu-Ramen.	Statement + Select + Shop
0037	O	はいそれでは金龍ラーメンへご案内いたします Then, I'll guide you to Kinryu-Ramen.	Exhibit + Guide + Shop
0038	D	はい Thank.you.	Statement + Exhibit + YesorNo

Fig. 4. Example of a dialogue with LIT

3.1. Speech Intention Annotation

For the robust understanding of in-car spoken dialogues, we designed layered intention tag [6]. The intention tag expresses the task-dependent intention of speaker, such as request of search, statement of the search result and so on. The layered intention tag (LIT) is composed from the four layers, "Discourse act", "Action", "Object" and "Argument". Figure 3 shows a part of the organization of LIT. As Figure 3 shows, the lower layered intention tag depends on the upper layered one.

We have tagged for over 35,000 utterance units manually and built the spoken dialogue corpus with LIT which is composed of 3641 conversations in 1256 sessions (Table 1 (a)). Figure 4 shows a sample of restaurant guide dialogue with LIT. There exist 95 types of LIT in the corpus.

To evaluate the reliability of LIT, we made the experiments by several annotators and evaluated by Cohen's kappa value. As the results of the experiments we confirmed the

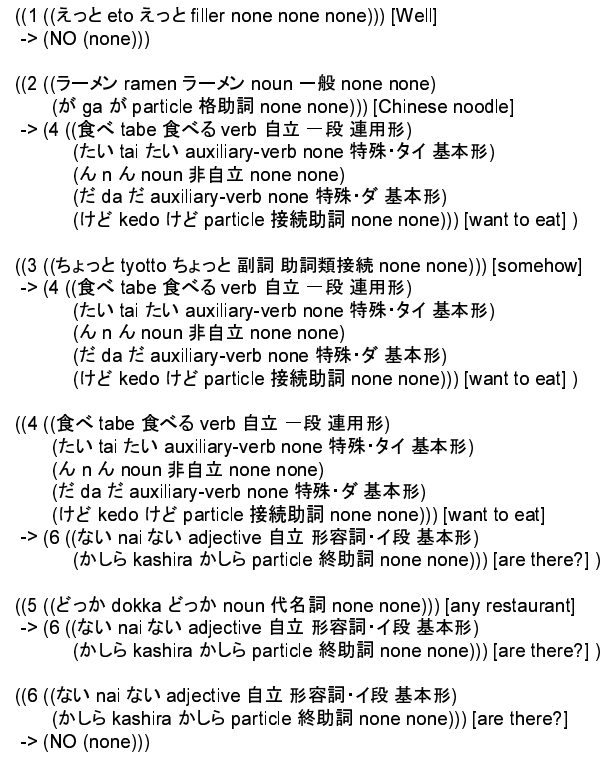


Fig. 5. Dependency structure of No.0028 utterance

reliable corpus data was built.

3.2. Dependency Structure Annotation

To characterize spontaneous dialogue speeches from the viewpoint of dependency, we have constructed a syntactically annotated spoken language corpus by providing morphological and syntactic information for each of the driver's utterances in CIAIR in-car speech dialogue corpus [7].

We have provided boundaries between words, pronunciation, basic form, part-of-speech, conjugation type and conjugated form of each word as the morphological information, and boundaries and dependencies between bunsetsus [1] as the syntactical information. In Japanese a dependency is a modification relation that a dependent bunsetsu depends on a head bunsetsu. That is, a dependent bunsetsu and a head bunsetsu work as a modifier and a modifyee, respectively.

Figure 5 shows the dependency structure of No. 0028 utterance in Figure 4. It illustrates a sequence of dependency relations, each of which consists of a dependent bunsetsu and a head bunsetsu. Each bunsetsu is listed with its number and its constituent morphemes.

[1] A *bunsetsu* is one of the linguistic units in Japanese, and roughly corresponds to a basic phrase in English. A bunsetsu consists of one independent word and more than zero ancillary words.

Table 2. POD (Part-Of-Dialogue) and its role

POD	Role
guide	guiding to restaurant or parking
srch	searching a restaurant
p_srch	searching a parking
slct	selecting restaurant or parking
genre	choosing a cuisine
rsrv	making reservation
srch_rqst	requesting search
rsrv_rqst	requesting reservation
s_info	extracting shop information such as menu, price, reservation, area.
p_info	extracting parking information such as vacant space, near alternatives.
rsrv_dtl	extracting reservation information such as number of people, time.

We have annotated the corpus by providing morphological and syntactic information for about 14,000 utterances. In the corpus, there are over 85,000 morphemes and over 45,000 dependency relations (Table 1 (b)).

3.3. Dialogue Structure Annotation

To represent the structure of dialogue, we consider a dialogue as a sequence of LIT and have provided structural trees [8]. If the dialogue-structural rules are obtained from the structural trees, it would be able to apply usual language parsing technologies to the analysis of the dialogues. In this research, we used from 1st to 3rd layers of LIT and extended LIT with speaker symbol like "D+Request+Search+Shop".

To construct a structural tree, we defined the category called POD (Part-Of-Dialogue), according to the observation of the restaurant guide dialogue. Table 2 shows 11 types of POD and the role which the POD plays in the dialogues. The POD represents a partial structure of the dialogue.

We provided structural trees for 789 dialogues and obtained 297 dialogue-structural rules (Table 1 (c)). Figure 6 shows the dialogue-structural tree of the dialogue of Figure 4. In Figure 6 the LIT is represented in initials like DRSS. Because we express the dialogue structure of restaurant guide dialogue, the root of the dialogue-structural tree is "guide" POD.

4. ANALYSIS AND UTILIZATION OF IN-CAR SPOKEN DIALOGUE

4.1. Speech Intention Understanding using Decision Tree Learning

To determine the intention of an utterance, we constructed 32 decision trees [9] by using the intention-annotated spo-

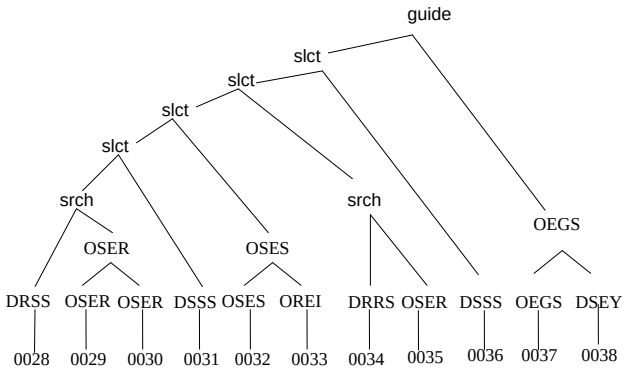


Fig. 6. Dialogue-structural tree for the dialogue of Figure 4

ken dialogue corpus with LIT.

As a set of attributes, we used the present speaker, the previous LIT (from 1st to 3rd layers), the previous speaker and the morphemes appearance.

32 decision trees are shown in Figure 7. The inference algorithm using these decision trees is as follows,

1. The four trees in 1st group are used and one tree which has the lowest re-classification error rate (the lowest error rate in training data) is chosen.

2. The tag obtained in first step, for example "Action" layer, is added to the attribute set and the three trees in II of 2nd group are used. Then, one of the three trees is chosen as same as first step.

3. The tag obtained in second step, for example "Discourse act" layer, is added to attribute set and the two trees in A of 3rd group are used.

4. One tree in 4th group is used for inference of the remaining undecided layer.

In our experiment, 1218 dialogues in the corpus with LIT (Section 3.1) were used. We divided 3143 driver's utterances into two groups. One is the training data which consists of 2972 utterances, the other is the evaluation which consists of 171 utterances. We used See5[2] for decision tree learning. As a result of evaluation experiment, the detection precision is 73.1%.

4.2. Speaker's Knowledge Acquisition from Dialogue Data

In the restaurant guide dialogues, various restaurant informations might be included. To extract the restaurant information from the dialogue, we utilize dependency structures of utterances. Using the dependence structure annotation tool we produced the dependency structure of each

[2]http://www.rulequest.com/see5-info.html

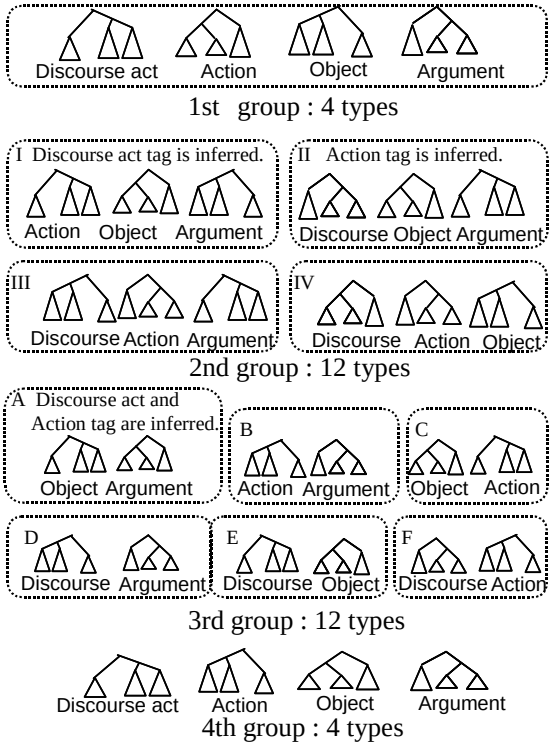


Fig. 7. 32 decision trees

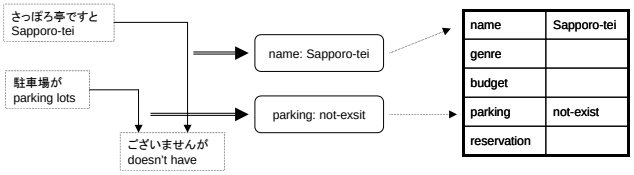
Table 3. Items extracting from restaurant guide dialogues

item	content
name	the name of a restaurant
genre	the kind of cuisine
budget	the estimation for the meal
parking	existence of parking lots
reservation	necessity or possibility

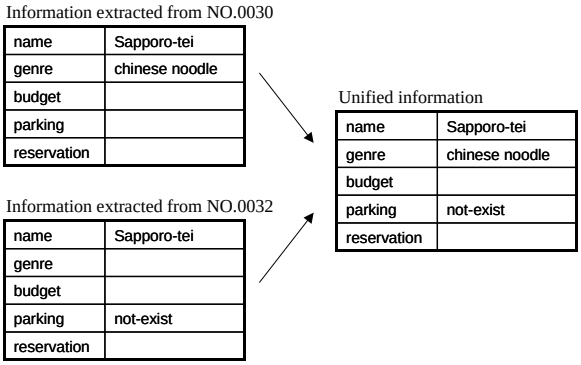
utterance in the dialogue. In this method we extracted the restaurant information from each utterance in the dialogue and unified them through the whole dialogue. The items of the restaurant information are listed in Table 3.

Figure 8(a) shows the extraction from No.0032 utterance in Figure 4 and Figure 8(b) shows the unification of No.0030 & No.0032.

To evaluate the availability of the method we did the evaluation experiment using 100 restaurant guide dialogues (937 utterances) which have 702 restaurants' information. As the result, the precision is 83.3% and the recall is 64.1%.



(a)Information extraction from No.0032 utterance



(b) Unification of No.0030 & No.0032

Fig. 8. Information extraction and unification

5. DEVELOPMENT OF SPOKEN DIALOGUE SYSTEMS

5.1. Corpus-Based Spoken Dialogue System

We constructed a prototype spoken dialogue system as a workbench for the evaluation of the technologies for different part of spoken dialogue system [10]. Figure 9 shows the system configuration and the information flow between components.

Most of the components in the system are implemented by using the statistical information obtained from the corpus. The speech understanding calculates the similarity between the input sentence and the example whose previous intention tag is the same as the tag of the preceding system response [11]. In order to decide an intention of system's response, we use 3-gram of the intention tags created from the corpus with LIT.

From the result of the evaluation experiment with 6 subjects, the rate of task completion have improved 20% by adding the examples data.

As a future work, we consider applying the speech intention understanding using the decision trees (Section 4.1) and the speech management using the dialogue-structural trees (Section 3.3) to the spoken dialogue system.

5.2. Example-Based Spoken Dialogue System

We have proposed a spoken dialogue control technique using dialogue examples and a technique to collect dialogue

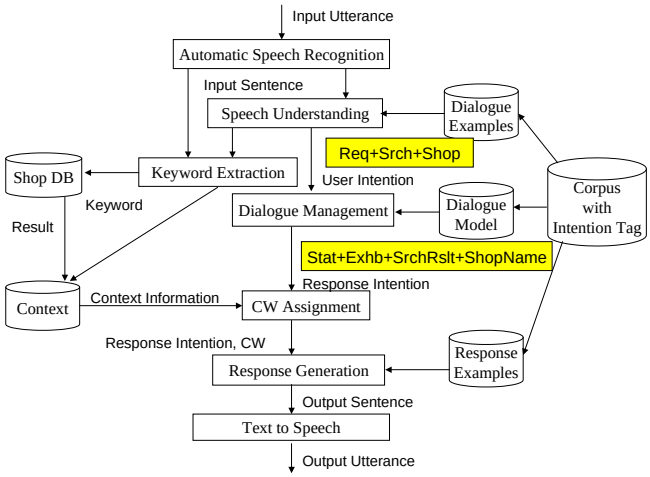


Fig. 9. Configuration of a prototype spoken dialogue system

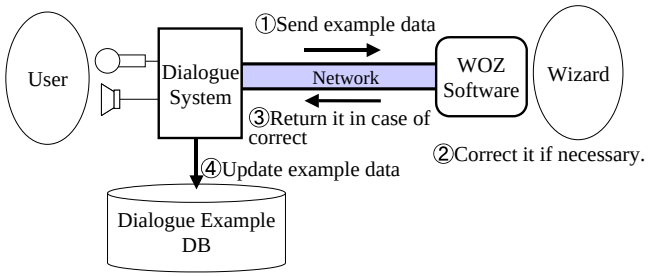


Fig. 10. Configuration of the "GROW" architecture

example data from the dialogue performed between a human subject and a pseudo-spoken-dialogue system based on WOZ scheme [12]. And we extended the technique to be able to handle context dependent utterances [13].

This architecture is named "GROW". As Figure 10 shows, the dialogue system and the WOZ system are connected with network. The reply created by the dialogue system are transferred to the WOZ system, the Wizard corrects it only if it is necessary. The data which is corrected by the Wizard is sent to the dialogue system, presented to the user and preserved in the dialogue example data base. When the correction is unnecessary, the reply which is generated by the dialogue system is presented to the user as it is. In this architecture, the system can automatically add the correct dialogue data as a new example.

As a result of an evaluation experiment with human subjects, the success rate of reply generation has improved as the number of the examples increased.

6. CONCLUSION

We have introduced our research activities, and presented the various fruits of the in-car speech corpus. All these researches were accomplished only by utilizing a large-scale corpus. Generally, the conversation in car is task-oriented and simpler than the general conversation. Therefore, the in-car spoken dialogue corpus is really useful for the spoken dialogue system research.

In the future, we will keep researching using the corpus, the tool, and the prototype system which are introduced in this paper.

7. ACKNOWLEDGMENTS

This work is partially supported by a Grant-in-Aid for COE Research (No. 11CE2005) and for Scientific Research (No. 15300045) of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

8. REFERENCES

[1] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proceedings of 2nd International Conference on Language Resources and Evaluation*, 2000, pp. 947–952.

[2] N. Kawaguchi, S. Matsubara, K. Takeda, and F. Itakura, "Multi-dimensional data acquisition for integrated acoustic information research," in *Proceedings of 3rd International Conference on Language Resources and Evaluation*, 2002, pp. 2043–2046.

[3] N. Kawaguchi, S. Matsubara, I. Kishida, Y. Irie, Y. Yamaguchi, K. Takeda, and F. Itakura, "Construction and analysis of the multi-layered in-car spoken dialogue corpus," in *Proceedings of the Workshop on DSP in Mobile and Vehicular Systems*, 2003.

[4] N. Kawaguchi, K. Takeda, and F.Itakura, "Multimedia corpus of in-car speech communication," *The Journal of VLSI Signal Processing*, vol. 36, no. 2, pp. 153–159, 2004.

[5] N. Kawaguchi, S. Matsubara, K.Takeda, and F. Itakura, "CIAIR in-car speech corpus -influence of driving states-," *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, pp. 578–582, 2005.

[6] Y. Irie, S. Matsubara, N. Kawaguchi, Y. Yamaguchi, and Y. Inagaki, "Design and evaluation of layered intention tag for in-car speech corpus," in *Proceedings of International Symposium on Speech Technology and Processing Systems*, 2004, pp. 82–86.

[7] T. Ohno, S. Matsubara, N. Kawaguchi, and Y. Inagaki, "Robust dependency parsing of spontaneous japanese spoken language," *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, pp. 545–552, 2005.

[8] S. Kato, S. Matsubara, Y. Yamaguchi, and N. Kawaguchi, "Construction of structurally annotated spoken dialogue corpus," in *Proceedings of 5th Workshop on Asian Language Resources*, 2005 (in printing).

[9] Y. Irie, S. Matsubara, N. Kawaguchi, Y. Yamaguchi, and Y. Inagaki, "Speech intention understanding based on decision tree learning," in *Proceedings of 8th International Conference on Spoken Language Processing*, 2004.

[10] K. Hayashi, Y. Irie, Y. Yamaguchi, S. Matsubara, and N. Kawaguchi, "Speech understanding, dialogue management and response generation in corpus-based spoken dialogue system," in *Proceedings of 8th International Conference on Spoken Language Processing*, 2004.

[11] S. Matsubara, S. Kimura, N. Kawaguchi, Y. Yamaguchi, and Y. Inagaki, "Example-based speech intention understanding and its application to in-car spoken dialogue system," in *Proceedings of 19th International Conference on Computational Linguistics*, 2002, pp. 633–639.

[12] H. Murao, N. Kawaguchi, S. Matsubara, Y. Yamaguchi, and Y. Inagaki, "Example-based spoken dialogue system using woz system log," in *Proceedings of SIGdial Workshop on Discourse and Dialogue*, 2003, pp. 140–148.

[13] H. Murao, N. Kawaguchi, S. Matsubara, Y. Yamaguchi, and Y. Inagaki, "Example-based spoken dialogue system with online example augmentation," in *Proceedings of 8th International Conference on Spoken Language Processing*, 2004.