社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006−SLP−61 (6)
2006／5／11

対話コーパスに基づく階層化された発話意図の推定

入江友紀[†1,†5] 松原茂樹[†2] 河口信夫[†3]
山口由紀子[†2] 稲垣康善[†4]

本論文では、機械学習に基づく発話意図推定の一手法を提案する。本研究では、対話システムにおいて意図推定の結果を直接的に利用することを目的に、階層化された発話意図を用いる。これは、話者の意図に関わる情報を、その詳細度に応じて体系化したものである。本手法では、多様なレベルの意図情報を正しく推定するために、機械学習により獲得した決定木を意図推定ルールとして使用する。発話意図が階層化されていることを考慮し、階層ごとに決定木を作成し、それらを動的な順序で適用する。名古屋大学 CIAIR 車内音声対話コーパスを用いて意図推定実験を実施した。2972 発話から学習した決定木による発話意図推定の正解率は 73.1%であり、本手法の有効性を確認した。

Layered Speech-Act Detection based on Spoken Dialogue Corpus

Yuki Irie,[†1,†5] Shigeki Matsubara,[†2] Nobuo Kawaguchi,[†3]
Yukiko Yamaguchi[†1] and Yasuyoshi Inagaki[†1]

This paper proposes a technique for speech intention understanding based on decision tree learning. The technique can extract the feature quantity strongly related to the intentions, and therefore cope with the addition of dialogue examples robustly. The organization of intention tag has been designed hierarchically according to the informational degree of details, and the tag more detailed than an illocutionary act is given to each utterance unit. In speech intention detection, high accuracy can be realized by adopting the procedure of processing sequentially from the decidable class. An experiment on driver's speech intention detection has been made using the restaurant search conversations of an in-car spoken dialogue corpus. The decision trees for intention detection were created from 2,972 utterance units to which the intention tags were given. As a result, 73.1 % of precision was acquired and the technique has been confirmed to be effective.

1. Introduction

In recent years, much of the progress in spoken dialogue systems has been made thanks to advances in speech recognition technology. In order to interact with a user naturally and smoothly, it is necessary for a spoken dialogue system to understand the intention of the user exactly.

A spoken dialogue system has to deal with utterances which include speech errors and fillers, or short utterances. However, it is difficult to understand the intention of such extra-grammatical utterances using a traditional AI approach where experts make rules and knowledge manually. On the other hand, a corpus-based approach to intention understanding could deal with such utterances robustly. Against the backdrop of collection of large-scale corpora[6),8)], many researches on speech intention understanding have been conducted so far[13)].

The approach of considering a dialogue corpus as case examples has been proposed. An example-based approach[9),10)] and machine learning are suitable for problem-solving based on case examples. An example-based approach (also called case-based learning[5)], and instance-based learning[1)]) uses a spoken dialogue corpus with dialogue act tags to regard the intention of each input utterance as being that of the most similar utterance in the corpus. Midgley[10)] calculated the degree of similarity according to speaker change, word number, word similarity, n-gram similarity, and previous and 2-previous DA tags. Matsubara et al.[9)] calculated the degree of similarity based on the degree of correspondence in morphemes and dependencies between utterances. The degree of similarity is weighted by the dialogue context information.

However, it is difficult to determine the intention of an utterance by an example-based approach when a similar example does not exist in the corpus. For that reason, many examples have to be stored in order to raise the precision of the classification. As the number of examples increases, the cost of similarity

†1 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University
†2 名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University
†3 名古屋大学大学院工学研究科電子情報システム専攻
Department of Electrical Engineering and Computer Science, Nagoya University
†4 愛知県立大学情報科学部
Information Technology Center, Nagoya University
†5 現在、株式会社デンソー
Currently, DENSO CORPORATION

calculation becomes high. Machine learning can efficiently make rules from relatively little data, and deal with new examples robustly. This also has the advantage that one can give the system a large number of potentially valuable features and allow it to determine which of the features are actually useful for identifying dialogue acts. With the recent availability of dialogue corpora, researchers have been investigating a variety of different machine learning techniques, such as Transformation-Based Learning[12], decision tree learning, for calculating the dialogue act of an utterance from features of the utterance itself and also from the dialogue acts of the preceding utterances.

In this paper, rules are made using machine learning. To make rules, we have used decision tree learning. The method constructs several decision trees for intention understanding. By constructing several decision trees, the characteristic features related to intentions can be retrieved, and it is also possible to deal with the diversity of the utterances robustly.

So far, we have designed a tag organization method known as layered intention tag (LIT). These tags reveal a more detailed utterance intention than the illocutionary act, and we have successfully built a corpus with this method[2]. Compared with the tags used for conventional corpus annotation, LIT is specialized for spoken dialogue systems. LIT is composed from 4 layers, taking into account the relevance between the intention and various phenomena relevant to an utterance, such as style, keywords and sentence structure.

Our method constructs 32 decision trees by using this corpus and detects the intention by combining them. Because decision trees are constructed for every LIT layer, the method can focus attention on those characteristics related to the intention of the layer. In order to evaluate the effectiveness of our method, we performed experiments for speech intention understanding. These experiments used the driver utterances about a restaurant search recorded on a large-scale in-car spoken dialogue corpus from CIAIR[3]. Decision trees were used, which were obtained from 2,972 utterance units in an intention-tagged corpus. As a result, the effectiveness of the method was confirmed.

2. Dialogue Corpus with Intention Tags

First, this paper gives a short sketch of the dialogue corpus which we have previously assembled using LIT.

2.1 CIAIR In-car Spoken Dialogue Database

In-car dialogue speech was recorded in a simulated running car environment in which some tasks are assigned to be performed in the car, such as a store search and guidance. The transcription of dialogue speech was based on the transcription criteria for the Corpus for Spontaneous Japanese (CSJ)[7].

The speech files are transcribed into ASCII text files by hand. As an advance analysis, linguistic phenomenon tags are assigned to fillers, hesitations, and so on. Furthermore, each speech is segmented into utterance units by a pause,

```
0003 - 00:04:955 - 00:06:560 M:D:N:O:
じゃあ                                    &ジャー
マック                    [McDonald's]    &マック
教えてください<SB>        [Please tell me] &オシエテクダサイ<SB>

0004 - 00:08:101 - 00:09:952 F:O:N:I:
はい                      [Yes]           &ハイ
マクドナルドですね<SB>    [McDonald's]    &マクドナルドデスネ<SB>

0005 - 00:10:665 - 00:14:111 F:O:N:O:
この先                    [Around here]   &コノサキ
二百メートル先に          [200 meters away form here] &ニヒャクメートルサキニ
マクドナルドが            [McDonald's]    &マクドナルドガ
あります<SB>              [There is]      &アリマス<SB>
```

図 1 Sample of transcription of dialogue speech

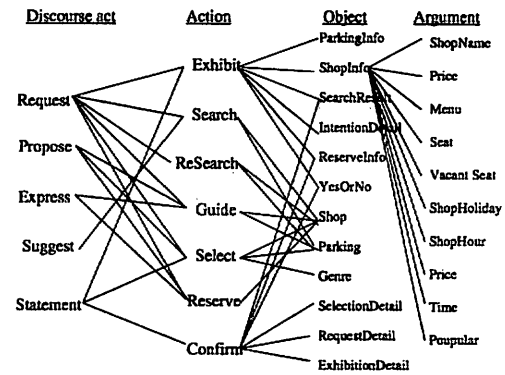

図 2 A part of the organization of intention tags

and the exact start time and end time are provided for them. Environmental information on sex (male/female), speaker's role (driver/navigator), dialogue task (navigation/information retrieval/...), and noise (noisy/clean) is provided for each utterance unit. An example of a transcript is shown in Fig. 1.

2.2 Design of Intention Tags

LIT[2] gives a detailed description of the speaker's intention to a task-dependent level *. Therefore, the intention ranges from a discourse act to detailed information such as "store name" and "price". Consequently, the tag was divided into layers by considering utterance elements strongly related with an intention. Fig. 2 shows a part of the organization of intention tags. The "Discourse act" layer denotes the role of the utterance unit in the dialogue. The "Action" layer denotes the action of the utterance unit. The "Object" layer denotes the object of the action such as "Shop", "Parking", etc. The "Argument" layer denotes other miscellaneous information about the utterance unit. Most of the argument layer tags can be decided directly from specific keywords in the sentence. As Fig. 2 shows, the upper-layer intention tag and the lower-layer one depends on each other.

* In this research, we have targeted at restaurant search task.

表 1 Example of an intention-tagged corpus

Transcription		Layered Intention Tag			
Speaker	Utterance unit	Discourse act	Action	Object	Argument
Driver	chu'uka no omise aru kana. (I'm looking for a Chinese restaurant.)	Request	Search	Shop	
Operator	chikaku ni "Daikokuten" ga ari masu. (There are "Daikokuten" near here.)	Statement	Exhibit	SearchResult	ShopName
Driver	sono mise ni rahmen ha aru no kana. (Does it serve Chinese noodles?)	Request	Exhibit	ShopInfo	Menu
Operator	hai gozai masu. (Yes, it does.)	Statement	Exhibit	ShopInfo	Menu
Driver	soko ni an'nai site. (Please guide me there.)	Request	Guide	Shop	
Operator	"Daikokuten" made goan'nai shimasu. (Ok. I'll navigate to "Daikokuten".)	Express	Guide	Shop	

2.3 Building an Intention-tagged Corpus

For building an intention-tagged corpus, we have used the CIAIR transcribed corpus. For each utterance unit about restaurant search in the corpus, we provided the intention tag by hand[2]. At this time, we have tagged over 35,000 utterance units. When we built the transcribed corpus, each utterance was divided into utterance units by pauses of 200 ms or more. Therefore, one intention tag is given to one utterance unit in principle. Table 1 shows an example of an intention-tagged corpus.

3. Speech Intention Understanding based on Decision Tree Learning

3.1 Overview of the Method

The method constructs decision trees using an intention-tagged corpus. In this research, the software See5 ** was used for decision tree learning. If we input a set of data (an attribute and its value, and a class), this software outputs a decision tree. As an attribute set, the present speaker, the previous intention tag (discourse act layer tag, action layer tag and object layer tag), the previous speaker and morphemes appearing in training data are used. Because they were effective attributes in a preliminary experiment on speech intention understanding.

LIT, which is a detection object in this research, is composed from 4 layers, which are strongly dependent on each other. The discourse act layer would be related to the sentence end (modality) and the argument layer would be related to keywords. Therefore, effective features would differ from layer to layer.

The element which is strongly related with an intention may differ from LIT layers or utterances. When an intention-tagged corpus was built, annotators took advantage of this information and tagged from a layer where the tag is easy to determine. Also, they narrowed down candidate tags for other layer by using restrictions on layers and the already determined tags.

In our study, in order to understand speech intention flexibly, decision trees are constructed for every LIT layer. By considering elements strongly connected with a certain LIT layer and ordering decision trees for every utterance, the method can rapidly detect LIT layer which is the most strongly related to the utterance.

** See5: http://www.rulequest.com/see5-info.html

表 3 Classes and number of types

class	number of classes
Discourse act tag	5
Action tag	7
Object tag	12
Argument tag	36

Dialogue corpus with intention tags

Utterance	Intention tag
D: Does the shop serve Chinese noodles?	Request+Exhibit+ShopInfo+Menu

Data file

speaker	previous intention tag	previous speaker	morpheme: this	morpheme: the	morpheme: shop	morpheme: ...	tag already obtained	Class
D	Express+Exhibit+SearchResult	O	no	yes	yes	...	1-Request	2-Exhibit
D	Express+Exhibit+SearchResult	O	no	yes	yes	...	1-Request	3-ShopInfo
D	Express+Exhibit+SearchResult	O	no	yes	yes	...	1-Request	4-Menu

図 3 Example of data corresponding to utterance "Does the shop serve Chinese noodles?"(2nd group)

3.2 Learning Decision Trees

In addition to the attributes shown in Section 3.1 (present speaker, previous intention tag, previous speaker, and morphemes appearing in training data), the intention tag already determined is also used as an attribute. Since restrictions exist among layers as described in Section 2.2, tags for a certain layer would be dependent on tags for other layers. Tags for each layer were given as a class. Table 2 shows attributes and attribute values, and Table 3 shows classes and their number of types. Moreover, an example of the data corresponding to the utterance "Does the shop serve Chinese noodles?" is shown in Fig. 3.

Thirty-two types of decision trees were constructed by the data set obtained from an intention-tagged corpus. Fig. 4 shows an overview of the constructed decision trees.

1st group: when no tag of any layer is detected.

- Attribute

表 2 Attributes and attribute values

attribute	attribute value	
speaker	driver(D), operator(O)	
previous intention tag	"Statement2+Exhibit+ShopInfo" etc. 48 types	
previous speaker	driver(D), operator(O)	
morpheme	the morpheme appears in the utterance unit or not	
tags already obtained	Discourse act	5 types
	Action	7 types
	Object	12 types
	Argument	36 types

the present speaker, the previous intention tag, the previous speaker, morphemes appears in the utterance or not

- Number of decision trees
  4 types of decision trees (the decision tree for the discourse act layer, that for the action layer, that for the object layer and that for the argument layer)

2nd group: when the tag for one layer is detected.

- Attribute
  the present speaker, the previous intention tag, the previous speaker, morphemes appears in the utterance or not, tags already obtained
- Number of decision trees
  12 types of decision trees (Three patterns of the detected layer exist for each layer.)

3rd group: when the tags for two layers are detected.

- Attribute
  the present speaker, the previous intention tag, the previous speaker, morphemes appears in the utterance or not, tags already obtained
- Number of decision trees
  12 types of decision trees (Three patterns of the detected layer exist for each layer.)

4th group: when the tags for three layers are detected.

- Attribute
  the present speaker, the previous intention tag, the previous speaker, morphemes appears in the utterance or not, tags already obtained
- Number of decision trees
  4 types of decision trees (One patterns of the detected layer exist for each layer.)

3.3 Detection Algorithm

Tags for each layer are detected using decision trees obtained in Section 3.2. The order of the detected layer is determined by decision tree choice rules. These rules were defined as follows.

Decision tree choice rules

(1) The decision tree with the leaf which is the lowest rate of re-classification error (rate of error in training data) is chosen.

(2) When the rate of re-classification error is the same, the decision tree which had more training dates in the reached leaf is chosen.

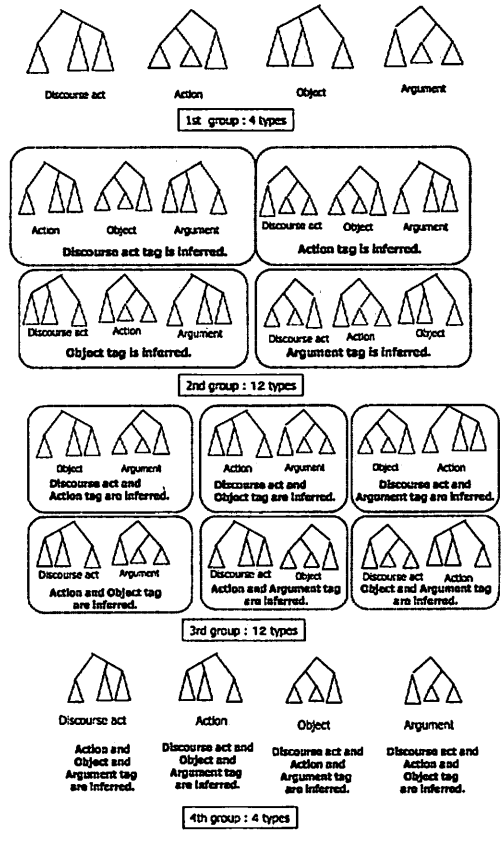

図 4 Overview of constructed decision trees

表 4 Detection precision (training data: 2,972 utterances, closed test)

tag	Discourse act	Action	Object	Argument
precision (%)	92.2	76.8	80.6	85.8

(3) When a decision tree cannot be chosen by the above rules, it is chosen in the following order: discourse act, argument, object, and action tag. This is based on the result of a preliminary experiment which was executed on See5 using the learning data. The result is shown in Table 4.

According to the following algorithm, an utterance in-

tention may be detected.

Detection algorithm

(1) The candidate tags for each LIT layer are output using the four types of decision trees of the 1st group.
(2) One decision tree is chosen according to the choice rules, and the reached leaf in the chosen decision tree is regarded as a detected tag.
(3) The tag obtained by (2) is added to the attribute values, and the undecided layer tag is detected using the decision trees of the 2nd group.
(4) The tag obtained by (3) is further added to the attribute values, and the undecided layer tag is detected using the decision trees of the 3rd group.
(5) The tag obtained by (4) is further added to the attribute values, and the undecided layer tag is detected using the decision trees of the 4th group.

3.4 Example of Intention Detection

Fig. 5 shows an example of speech intention understanding of "How much is it?". The number (E/N) on leaves in Fig. 5 means the rate of re-classification error. N is the sum of the partial examples which reached a leaf, and E is the number of examples which belong to the class except the selected classes.

First, the candidate tags for each LIT layer is output using the four types of decision trees of the 1st group. In Fig. 5, "Request" is output as a candidate for the discourse act tag, "Exhibit" is output for the action tag, "ParkingInfo" is output for the object tag and "Price" is output for the argument tag. The discourse act tag "Request" is regarded as a detected tag, because its rate of re-classification error is the lowest (decision tree choice rule 1). Next, the "Request" tag is added to the attribute values, and the undecided layer tag is detected. The candidates for the undecided layer are output using the decision trees of the 2nd group. Then, "Exhibit" is output as a candidate for the action tag, "ShopInfo" is output as one of the object tags and "Price" is output as one of the argument tags. The "Price" is regarded as a detected result (decision tree choice rule 3). Similarly, the argument tag is added to the attribute values, and the undecided layer tag is detected using the decision trees of the 3rd group. Then, "Exhibit" is regarded as a detected result. Finally, the object tag is added to the attribute values and "ShopInfo" is detected using the decision trees of the 4th group. As a result, the system can arrive at a correct utterance intention "Request+Exhibit+ShopInfo+Price".

4. Experiments

In order to evaluate the effectiveness of the intention understanding method using several decision trees, we performed experiments using an intention-tagged corpus described in Section 2.

4.1 Outline of Experiment

In our experiment, 1,218 dialogues in an intention-tagged corpus were used, and Chasen *** was used for the morphological analysis. Nouns and proper nouns, which are characteristic of dialogue tasks, and filler, which is peculiar to spoken language, were added to Chasen's dictionary. The word class database was used for proper nouns such as restaurant names, menu[11].

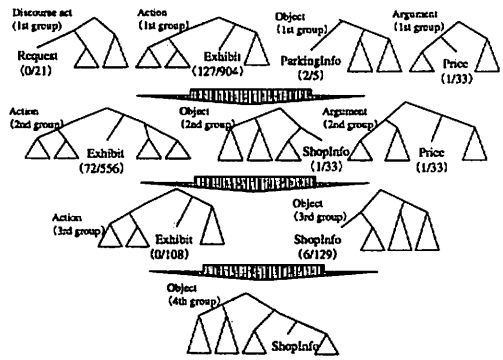

図 5 Flow of intention detection processing

We have divided 3,143 driver's utterance units into two groups. One is training data of 2,972 utterance units (183 persons), and the other is test data of 171 utterance units (7 persons). There exists no duplication between them. 41 kinds of the dialogue tasks were prepared for collection of restaurant search conversations in the CIAIR in-car speech corpus, and the test data includes 18 kinds of them by chance. The decision trees were constructed by the decision tree learning software See5, and 1,371 kinds of morphemes appeared in the training data.

In our experiment, we determined a pre-annotated tag as the correct tag and checked the precision (the ratio of utterances which output the correct intention tag to total utterances). The following three patterns of decision trees were made, and their results were compared.

(1) Intention detection of all layers together
(2) Intention detection of each layer independently
(3) Intention detection using several decision trees

Fig. 6 shows a part of a decision tree for method 1. Method 2 uses all trees in 1st group of Fig. 4 and detects the intention by combining results for each layer.

4.2 Experimental Results

The experimental results are shown in Table 5. A precision of 67.8% was obtained by detection based on the above method 1, a precision of 56.7% using method 2 and a precision of 73.1% using method 3. This result shows our method using several decision trees to be more effective

Many sentence endings such as "onegai. (please.)", "...shite hoshi'i. (I want you to...)", appear on the decision tree for the discourse act layer. Morpheme expressed system operations such as "an'nai (guide)",

*** Chasen: http://chasen.aist-nara.ac.jp/

表 5 Experimental result

	Intention detection of all layers together	Intention detection of each layer independently	Intention detection using several decision trees
precision (%)	67.8(116/171)	56.7(97/171)	73.1(125/171)

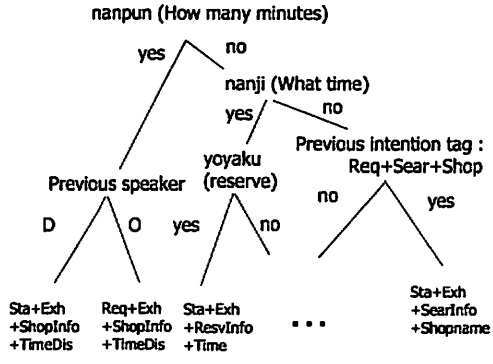

図 6 Example of a decision tree for method 1

"navigation", "yoyaku (reserve)" and the previous intention tag appear on the node of the decision tree for the action layer. In the decision tree for the argument layer, morphemes expressing time or price, etc. such as "nanpun (How many minutes)", "nanji (What time)", "ikura (How much)" appear. Thus, the elements strongly connected with each LIT layer differ and giving priority to these elements will improve precision.

Tags for other layers, having already been determined, appear on the host nodes. Because method 2 detects the tags independently, there are 48 mistakes (28%) due to the violation of tag restriction. Consequently, restrictions among layers may be utilized for speech intention understanding.

5. Concluding Remarks

This paper proposed a method of speech intention understanding based on spoken dialogue corpus. The method detects intention by constructing several decision trees and combining them. By constructing several decision trees, the characteristic features related to intentions can be retrieved, and it can also robustly cope with the diversity of the utterances. The experimental results, which use the decision trees obtained from 2,972 driver's utterance units in an intention-tagged corpus, show the effectiveness of our method.

In this paper, we have used morphemes as one of the attributes. Since the size of task vocabulary in this experiment was just 1,371, the corpus consisting of about 3,000 utterance units is not so small. However, if the vocabulary size becomes bigger, a large quantity of the corpus might be required. In experiments, it was effective for the proper nouns to classify them using the word class database. And therefore, it would be important to consider the classification of vocabulary, such as the usage of thesaurus.

参 考 文 献

1) D. W. Aha, D. Kibler and M. K. Albert: Instance-based Learning Algorithms, Machine Learning, Vol. 6, pp. 37–66 (1991).
2) 入江、松原、河口、山口、稲垣: 音声対話コーパスにおける発話意図タグの設計と評価, 電子情報通信学会論文誌, Vol. J88-D-II, No. 10, pp. 2169-2173 (2005).
3) N. Kawaguchi, S. Matsubara, K. Takeda and F. Itakura: CIAIR In-car Speech Corpus -Influence of Driving Status-, IEICE Trans. on Information and Systems, Vol. E88-D, No. 3, pp. 578-582 (2005).
4) S. Keizer, R. Akker and A. Nijholt: Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues, Proc. of 3rd SIGdial Workshop on Discourse and Dialogue, pp. 88–94 (2002).
5) J. Kolodner: An Introduction to Case-based Reasoning, Artificial Intelligence Review, Vol. 6, No. 1, pp. 3–34 (1992).
6) MADCOW: Multi-Site Data Collection for a Spontaneous Language Corpus, Proc. of ICSLP-92, pp. 903–906 (1992).
7) K. Maekawa, H. Koiso, S. Furui and H. Isahara: Spontaneous Speech Corpus of Japanese, Proc. of LREC-2000, pp. 947–952 (2000).
8) W. Marilyn, H. Lynette and A. John: Evaluation for DARPA Communicator Spoken Dialogue Systems, Proc. of LREC-2000 (2000).
9) 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, Proc. of COLING-2002, Vol. 1, pp. 633–639 (2002).
10) T. D. Midgley: Discourse Chunking: a Tool in Dialogue Act Tagging, ACL-03 Companion Volume, pp. 58–63 (2003).
11) H. Murao, N. Kawaguchi, S. Matsubara and Y. Inagaki: Example-based Query Generation for Spontaneous Speech, IEICE Trans. on Information and Systems, Vol. E87-D, No. 2, pp. 324–329 (2005).
12) K. Samuel, S. Carberry and K. Vijay-Shankerv: Dialogue Act Tagging with Transformation-Based Learning, Proc. of COLING/ACL-98, pp. 1150–1156 (1998).
13) A Stolcke, et al.: Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech, Computational Linguistics, Vol. 26, No. 3, pp. 339–373 (2000).